

Spatial Language and Dialogue

Edited by
Kenny R. Coventry, Thora Tenbrink, and John A. Bateman

03 Explorations in Language and Space

Spatial Language and Dialogue

EXPLORATIONS IN LANGUAGE AND SPACE

SERIES EDITOR Emile van der Zee, University of Lincoln

PUBLISHED

1 *Representing Direction in Language and Space*
Edited by Emile van der Zee and Jon Slack

2 *Functional Features in Language and Space: Insights from Perception, Categorization, and Development*
Edited by Laura A. Carlson and Emile van der Zee

IN PREPARATION

The Spatial Foundations of Cognition and Language
Edited by Kelly S. Mix, Linda B. Smith, and Michael Gasser

Spatial Language and Dialogue

Edited by

KENNY R. COVENTRY, THORA TENBRINK,
and JOHN BATEMAN



OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published in the United States
by Oxford University Press Inc., New York

First published 2009

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

Typeset by SPI Publisher Services, Pondicherry, India
Printed in Great Britain
on acid-free paper by
MPG Books Group

ISBN 978–0–19–955420–1

1 3 5 7 9 10 8 6 4 2

Contents

Preface

This book emerged from the *Workshop on Spatial Language and Dialogue*, organized at the Hanse Institute for Advanced Studies, Delmenhorst, Germany, in October 2005. The function of the workshop was to bring together researchers working in the fields of spatial language and dialogue in recognition of a distinct paucity of research in this area in spite of its obvious importance. We would like to thank the Engineering and Physical Sciences Research Council (grant no. EP/D033691/1 awarded to Kenny Coventry), the Hanse-Wissenschaftskolleg, and the DFG—SFB/TR8 Spatial Cognition Research Centre for jointly funding this workshop. We are most grateful to the programme committee, Laura Carlson, Christian Freksa, Simon Garrod, Christopher Habel, Michael Schober, Barbara Tversky, and Emile van der Zee, who all played an active role critiquing and selecting papers, and to additional reviewers who also provided helpful comments for each chapter. Also we are indebted to Wolfgang Stenzel and his colleagues at the Hanse-Wissenschaftskolleg, who facilitated the organization of the workshop, and to Emile van der Zee (series editor) and the editorial team at Oxford University Press for ensuring a smooth delivery of this volume. Finally, thanks go to David Smailes for assistance with the formatting. We hope you enjoy it!

Kenny Coventry
Thora Tenbrink
John Bateman

Notes on Contributors

John Bateman is Professor of English Applied Linguistics at the University of Bremen, Germany. His main research interests include natural language generation, functional linguistic approaches to multilingual and multimodal document design, ontology, dialogue systems and discourse structure.

Holly Branigan is Reader in Psychology at the University of Edinburgh, UK. Her main research interests are in language production and dialogue.

Laura Carlson is Professor of Psychology at the University of Notre Dame. She co-edited the volume *Functional Features in Language and Space: Insights from Perception, Categorization and Development* and is currently an associate editor of the journal *Memory and Cognition*.

Justine Cassell is Professor of Communication Studies and Electrical Engineering and Computer Science at Northwestern University, and the director of the Center for Technology and Social Behavior. Her interests concern the interaction between verbal and non-verbal behaviours in communication, and how these interactions can be embodied in conversational agents.

Kenny Coventry is Professor of Cognitive Science and Director of the Cognition and Communication Research Centre at Northumbria University, UK. His main research interest is the mapping between language and perception/action.

Marie-Paule Daniel is Maître de Conférences at the Université de Paris-Sud, Orsay. Her research focuses on navigation in realistic environments, most recently dysfunctions of spatial cognition in schizophrenic patients.

Anna Filipi is a Senior Research Fellow and Project Director at the Australian Council for Educational Research. Her research interests include bilingualism, language education, and the application of Conversation Analysis to first and second language acquisition, and to spatial language.

Julie Heiser is a Senior User Researcher at Adobe Systems Incorporated in San Jose, California. Her main research interests are spatial cognition and information representation as it relates to learning and memory.

Patrick Hill is a graduate student in the cognitive psychology programme at the University of Notre Dame.

Paul Lee is a Senior Research Associate for San Jose State University, working in the Human Systems Integration Division at NASA Ames Research Center. His main research interest is evaluating the roles of humans in the future air traffic system.

Martin Loetzsch is a PhD student at the Sony CSL in Paris working on the emergence of spatial language in autonomous robots. Before that he studied artificial intelligence and robotics and worked on robotic soccer.

Andrew Lovett is a doctoral candidate in the computer science programme at Northwestern University. He received his bachelor's degree in cognitive science from Yale University. Andrew is interested in developing computational models of cognitive processes and using these models to inform our understanding of the human mind.

Philippe Muller is Maître de Conférences in Computer Science at Toulouse University, France, and is working on natural language processing, mostly at the semantic level, specifically on the representation of temporal and spatial information.

Martin Pickering is Professor of the Psychology of Language and Communication at the University of Edinburgh, UK. His main research interests are in language production, language comprehension, dialogue, and bilingualism.

Laurent Prévot is a CNRS Postdoctoral Research Fellow in a linguistics team (CLLE-ERSS) at Toulouse University, France. His main research interest concerns discourse and dialogue organization from a multidisciplinary perspective.

Michael Schober is Dean and Professor of Psychology at the New School for Social Research, and editor of the journal *Discourse Processes*. He studies joint action in face-to-face and mediated settings, including in surveys with automated and virtual interviewers and among chamber musicians rehearsing and performing via remote video and audio.

Shi Hui is a senior researcher at DFKI-Lab Bremen, Safe & Secure Cognitive Systems and member of the DFG-funded SFB/TR 8 'Spatial Cognition' research centre. Among other topics she works in the following research areas: formal methods, shared-control, user modelling and dialogue modelling.

Timo Sowa is a software engineer at Elektrobit Corporation, Germany. His main research interests are gesture recognition and multimodal language comprehension from an interdisciplinary point of view.

Luc Steels is Professor of Artificial Intelligence at the University of Brussels (VUB) and Director of the Sony Computer Science Laboratory in Paris. His research interests cover the whole field of artificial intelligence and current work focuses on the foundations of semiotic dynamics and on fluid construction grammars.

Kristina Striegnitz is Assistant Professor in the Computer Science Department at Union College. Her research is on computational models for processing natural

language semantics and pragmatics, especially in the context of (multimodal) natural language generation and dialogue systems.

Thora Tenbrink is a postdoctoral researcher at the University of Bremen, Germany. She is a principal investigator in two projects concerned with the empirical investigation and interpretation of spatial language in natural discourse. Employing discourse analytic methods, she investigates linguistic reflections of cognitive principles underlying spatial and temporal language usage.

Paul Tepper is a PhD candidate in the technology and social behaviour joint PhD programme in computer science and communication studies at Northwestern University. In his current work, Paul is developing Embodied Conversational Agents (ECAs) that can build long-term social relationships with their users through conversation. He also works on computational models for generating coordinated language and iconic gestures.

Barbara Tversky is Professor Emerita of Psychology at Stanford University and Professor of Psychology at Columbia Teachers College. Her research includes memory, categorization, spatial cognition and language, event perception and cognition, and diagrammatic reasoning with applications to linguistics, computer science, human–computer interaction, education, and design.

Constanze Vorwerg teaches psycholinguistics at Bielefeld University, Germany. She is a researcher at the Collaborative Research Centre 'Alignment in Communication', and Member of the Scientific Board of the Cluster of Excellence 'Cognitive Interaction Technology'. Her research interests include spatial cognition, perceptual foundations of language, and alignment of situation models.

Ipke Wachsmuth is Director of the Center for Interdisciplinary Research (ZiF) and Chair of Artificial Intelligence at Bielefeld University, Germany. His main research interest is multimodal communication in embodied agents and intelligent virtual environments.

Roger Wales has recently retired as the Dean of Humanities and Social Science at La Trobe University, Australia. His research interests include psycholinguistics, spatio-cognitive interactions in the field of child language acquisition, and the effects of prosody in sentence processing.

Matthew Watson is a lecturer at the University of Sunderland, UK. His main research interest is the use of spatial language, in particular reference frames, in dialogue.

1

Spatial Language and Dialogue: Navigating the Domain

KENNY R. COVENTRY, THORA TENBRINK, and JOHN BATEMAN

1.1 Introduction

How one talks about where objects and places are in the world has occupied many researchers across diverse fields, such as linguistics, psychology, GIScience, architecture, and neuroscience. The popularity of this topic emanates from a realization that spatial relations play a pivotal role in language in a number of important ways. First, locating objects in the world is a rather basic facet of language that allows people to find objects and to direct the attention of others to relevant parts of the world. *The apple trees are behind the mountain* or *Look at the car over there* allow the hearer to find food or identify an object for joint attention. Spatial language therefore enables us to find objects and places outside the immediate stimulus reach of the speakers, making use of a wealth of spatial relations, spatial perspectives, and spatial objects, all finely tuned to the immediate needs and abilities of the interlocutors. Second, as many people have noted, there is something basic about space that is critical to structuring other aspects of our experience and language. Expressions such as *I'm over the moon* and *I'm really down* illustrate that positive and negative emotional states are often associated with high and low positions respectively, while expressions such as *I'm in the doldrums* or *It's beyond me* rely on spatial constructions of 'containment' and current 'mental' position. And third, at a level with direct practical applications, understanding spatial language is critical to improving how technology assists us in finding objects and locations outside the immediate observable range.

Developing technologies, such as GPS, now allow us to pinpoint with impressive accuracy the exact positions of people, places, and objects in the world. Ironically that doesn't mean that they are easy to find. Using language as a means of communicating such information affords a level of accessibility that formalizations of space often miss. Typically, implementations of spatial relationships in technical systems presuppose a fine level of metric information that is not reflected by human spatial thinking, and therefore remains entirely unrelated to natural communication. This complicates immensely the task of building suitable

interfaces between spatially-aware systems, such as Geographic Information Systems, situated robots, and their human users. To make progress here it is essential that we understand far more of how precisely humans deal with spatial information and communicate about that information. Indeed, the analysis of language as an accessible external representation reveals crucial facts about our internal conceptualizations. Such concepts represent qualitative, non-veridical models of the outside world, filtered and shaped by human perception and categorization processes. And so the study of spatial language and the study of our perception and conceptualization of space go hand in hand.

The most natural context of use for spatial language is in dialogic situations, where one interactant may need to know about the position of some object, or how to get somewhere, or just how various objects are to be brought together (for example, when being instructed how to construct something). Indeed talking about space can be regarded as a prototypical dialogue situation. Face-to-face communication is situated within space and time as shared common ground for the interactants, and therefore regularly relates to the current spatial setting in more than one way. We negotiate the location of objects, identify them by way of referring to their spatial position in relation to other objects, or give route directions to each other. All of these everyday discourse goals entail establishing common ground and coordinating the varying perspectives we can take on the same situation so that we develop a shared model of the world. It is no doubt because of this that spatial language exhibits such an astonishing degree of flexibility in the perspectives that it supports and the range of ways it offers for picking out particular spatial relationships and attributes.

Moreover, since most language occurs in dialogue situations, it is of central importance to identify just how participants align themselves with one another, find common reference frames, and identify the particular spatial properties and objects at issue when talking about the spatial world. There is growing recognition in the language research community at large that dialogue rather than monologue should be a starting point for language understanding (e.g. Clark, 1996; Pickering and Garrod, 2004). In many recent approaches, ranging from socio-semiotics (see Thibault, 2006), through neurocognition (e.g. van Berkum et al., 2008), to work in robot–robot and human–robot interaction (see Steels, 2003, for a recent review), concepts and language are seen as emergent properties of interaction between agents. Hence, the current Zeitgeist in both language research and robotics/AI demands an integrated examination of spatial language in dialogue settings. Furthermore, language comprehension, as well as conveying meaning, has a necessary social function. Being involved in an interaction is in itself as important as the transportation of information from one speaker to another.

Early work on spatial language certainly confirmed how crucial it was to consider the dialogic context when trying to explain the language options mobilized by speakers. For example, Garrod and Anderson (1987) examined how pairs of people direct each other to points on a maze over time, and found that dyads

quickly converge on shared ways of talking about spatial arrays. However, perhaps surprisingly from today's perspectives, and despite the fact that research on spatial language has been a hive of activity over the last few years (for example, authored books and edited collections include Carlson and van der Zee, 2005; Coventry and Garrod, 2004; Levinson, 2003; van der Zee and Slack, 2003), the vast majority of work in the field to date has examined spatial language in *monologue* situations, often in highly artificial and restricted settings.

As a direct response to this, we felt compelled to organize a workshop for researchers from the dialogue and spatial language communities to accumulate what we know about spatial language and dialogue, and to pinpoint what we do not know but need to research. This workshop, which took place at the Hanse Institute for Advanced Studies in October 2005, brought together leading researchers working on spatial language and on dialogue from a number of different perspectives spanning the full range of disciplines in the cognitive sciences. The result was a fertile ground for discussion of spatial language and dialogue, and we hope that this volume of selected papers emerging directly from the workshop conveys some of the excitement of that event. Furthermore this book can be treated both as a timely overview of the state of the art in this young field and as a call to arms for other researchers to focus on this important topic.

1.2 Organization of Volume

We have organized the order of contributions in this volume to move along several trajectories simultaneously, all crucially concerned with the relationship of spatial language and its mobilization in dialogue. The following guide should help readers in different disciplines to navigate their way through the volume and to draw connections between the contributions productively.

The initial chapters concern themselves with our basic starting point—the fact that spatial language and dialogue really do need to be seen together. Here we see several contributions that deal with methodological issues involved in this shift and some of the most pertinent models of language behaviour that are contributing to it. Matthew Watson, Martin Pickering, and Holly Branigan set this stage by examining the importance of looking at language within the context of dialogue, and overviewing the advances the literature on spatial dialogue has made in recent years. Specifically, they investigate the processes of interactive alignment influencing choices of spatial reference frames during dialogue.

The dialogic situation may have some striking effects on spatial language use, where language responds to the abilities of the interlocutors, to the dialogue history, and to the spatial context including both speakers as embodied spatial agents and spatial events. Thus, Michael Schober addresses spatial dialogue processes between partners with either high or low mental rotation abilities. His results highlight a number of systematic differences in speakers' descriptions based on

spatial ability. High-ability participants displayed better performance both in terms of choosing appropriate verbalizations and in terms of spatial perspective taking. They were able to subtly assess their interaction partner's ability quickly, and flexibly adjusted their utterances accordingly in suitable ways. Pairs with low-ability partners had substantial difficulties in interpreting each other's meaning.

Constanze Vorwerg then brings out a further critical benefit of placing spatial language in dialogue by focusing on the consistency in speakers' spatial descriptions across dialogue contributions with regard to lexical choices as well as reference systems. Her results show that speakers make their initial choices on the basis of the spatial relationship between reference object and target object in the first task, choosing a prototypical axis for a reference system. In subsequent descriptions, they tend to re-use the same conceptual and lexical patterns, basically self-aligning across trials in localization sequences. These results do not reflect default patterns or cognitive styles, since they could be systematically manipulated by the initial configurations providing good spatial relationships for different reference systems.

An essential aspect of dialogue is interaction, that is, how the to-and-fro of turn-taking with feedback from other interlocutors itself contributes significantly to the spatial language that is used. The next two chapters take up very different aspects of this role of interaction. Anna Filipi and Roger Wales draw on Conversation Analysis, and its reliance on understanding as an interactionally situated achievement, to investigate a Map Task scenario with respect to shifts in perspective as reflected by the deictic motion verbs *come* and *go*. Shifts in the usage patterns of these verbs align with shifts in spatial perspective. Regular patterns in perspective shifts relate to the speakers' shared information status as well as the linguistic and interactional signalling of task completion. Generally, however, shifts to a different perspective were avoided particularly by children, reflecting the additional cognitive load associated with changes in perspective.

The creation of spatial meaning in interaction is then taken to its extreme in the contribution of Luc Steels and Martin Loetzsch, who present an account of dialogic spatial communication between situated embodied agents that are specifically designed to model human dialogue processes. Steels and Loetzsch show how important features of natural spatial language arise out of the necessity of achieving successful interaction. This permits them to pinpoint ways in which perspective taking and marking comes into play in spatial communication in general. In particular, they focus on the necessity of taking perspective into account as soon as there is more than one point of view present, on the agents' ability to take another's perspective, and on the effects of explicitly marking perspective. If perspective is left implicit, interpretation is still possible by identifying possible matches, but involves higher cognitive costs.

In the next two chapters, we see work on the kinds of spatial descriptions that are produced in a variety of discourse situations. Such work is critical to obtain a more accurate view of the range of language that is actually produced by speakers

and to show how language reflects diverse communicative requirements, as well as how individual languages can lead to differing options being taken up in concrete situations.

Laura Carlson and Patrick Hill investigate speakers' productions of spatial descriptions across various discourse contexts in relation to an office scene. In theory, the selection of a spatial description may be guided either by choosing a particularly salient reference object, or by relying on a particularly prototypical spatial relation between target and reference object as expressed by the spatial term, or by a combination of both. Carlson and Hill's results show that the main factor influencing speakers' choices is the spatial relation; speakers prefer using a spatial term that refers to a position on a prototypical axis (consistent with Vorwerg's findings). Object saliency comes into play if there is more than one such relation, or if no good spatial relation is available in the scene.

Thora Tenbrink's chapter focuses on a comparison between English and German lexical choices in spatial descriptions. In her web-based empirical study, she has identified regular patterns in spatial object reference that systematically differed according to the language used. Although the underlying communicative principles were generally shared, the particular linguistic preferences depend on the specific repertory each language offers as well as a number of generalized preferences shared by each language community.

As mentioned in several of the contributions up to now, particularly when we move to increasingly natural interactions, it becomes necessary to consider other modes of spatial communication that regularly accompany spatial language and which augment and complement spatial language in important ways. Language as a communicative medium about space is not, after all, the only way to communicate about space, and indeed sometimes other means, such as gesture, can be more effective. The next few chapters explicitly bring this aspect of spatial communication to the fore.

Barbara Tversky, Julie Heiser, Paul Lee, and Marie-Paule Daniel begin by contrasting how participants perform two spatial tasks using different communicative modes (language, gesture, diagrams). Both tasks concern how people give explanations, but anchored in two contexts: how to get from A to B, and how to put something together. They show that different modes of explanation vary in their strengths and weaknesses dependent on task. For example, diagrams and gestures have the advantage of conveying information about actions with respect to objects in space directly, whereas language affords completeness and qualification. Nevertheless, Tversky *et al.* also show that explanations across these modes share common structures, such as similar semantics to convey actions at landmarks, and relate these to very general properties of communication that hold regardless of communicative mode. This is therefore an important step towards extending our view of what is considered when we talk of spatial dialogue.

Building particularly on analyses of gesture in spatial communication, in the next chapter Timo Sowa and Ipke Wachsmuth propose a classification of gestures

accompanying language that are used for communicating information about another spatial property: shape. Based on corpus data, they identify 84 distinct gesture kinds that are subsumed in four categories. Most gestures represent an object's outer dimensions, such as extent and profile. These are equally directly expressed by the associated linguistic expressions that the gestures regularly co-occur with; both of these types of information are given within a particular structural context. The authors propose a computational model implementing their findings, the Imagistic Description Tree, which serves to interpret multimodal shape-related expressions for the purposes of a gesture-understanding system.

The next chapter, that of Kristina Striegnitz, Paul Tepper, Andrew Lovett, and Justine Cassell, serves as a bridge to the final topic of the book, route instructions. Striegnitz *et al.* investigate how gestures accompanying verbal route directions indicate the location of landmarks, with a specific focus on how the underlying spatial perspective is reflected in gestures. Their empirical findings suggest, on the one hand, that route perspective is the preferred underlying conceptualization for landmark gestures and, on the other hand, that gestures sometimes reflect the speaker's actual position in relation to a landmark, rather than an underlying route or survey perspective. The authors propose a way of implementing these results in an embodied conversational agent, which is capable of providing route directions to humans by employing generated locating gestures together with language.

Although route instructions are certainly one of the most common types of spatial communication and have been mentioned in several of the preceding chapters, they have, like many other areas of spatial language, received study predominantly from the monologic perspective. Route instructions produced in dialogue show some crucial differences to route descriptions produced as monologues and bring in many of the detailed phenomena and issues that previous chapters have raised. The final two chapters focus in more detail on this type of spatial communication.

Philippe Muller and Laurent Prévot investigate an aspect of route descriptions that does not occur in monologic route description at all: feedback strategies. In their qualitative analysis of telephone conversations, they relate various types of dialogue acts and spatial content types to a range of feedback variants used for acknowledgement, grounding, anchoring, and the like. Their work highlights the role of different lexical cues in establishing mutual understanding within a spatial task. Its specific merit lies in the particularly naturalistic way of collecting dialogic data via phone calls, in which the participants were truly motivated to find a joint mental representation.

Finally, Shi Hui and Thora Tenbrink bring the book to a close with their investigation of route description dialogues between human users and a robotic wheelchair. Users' linguistic and conceptual representations may differ fundamentally and systematically from the robot's implemented functionalities and so raise issues of mismatched abilities, discussed by Schober, in a striking form. In order to deal with the communicative and interactive problems that are bound to arise, clarification dialogues must be initiated in order to align the spatial

representations in suitable ways. The authors use a Wizard-of-Oz study to identify a range of potential problem areas in relation to the robotic wheelchair's implemented conceptual route graph, and propose a dialogue model targeting fluent and intuitive discourse.

Altogether, the contributions in this book draw a diverse picture of current research within the area of spatial communication. At many points, it has become apparent that detailed and focused investigations of particularly *dialogic* situations still need to be undertaken in order to provide a broader basis for generalizable conclusions on spatial communication. This book offers a precise account of the state of the art concerning various endeavours in this direction, and we believe this shows just how fruitful and important future research in this area will be.

2

Why Dialogue Methods are Important for Investigating Spatial Language

MATTHEW E. WATSON, MARTIN J. PICKERING, and HOLLY P. BRANIGAN

2.1 Introduction

There is now a voluminous literature on the use of spatial language (e.g. Bloom, Peterson, Nadel, and Garrett, 1996; Carlson, 2003; Coventry and Garrod, 2004; Landau and Jackendoff, 1996; Levinson, 2003). However, it is notable that most experimental research has investigated spatial language in a monologue context. This is of course no different from other areas of psycholinguistics. An isolated context makes experimental control much more straightforward. To the extent that most psycholinguists have shown any interest in dialogue, they tend to assume that language in dialogue can be understood from studying language in monologue (see Clark, 1996; Pickering and Garrod, 2004). Under this assumption, understanding how people produce and comprehend utterances in isolation will lead to the understanding of how people communicate. However, there is more to dialogue than just the production of an utterance and its passive comprehension. Dialogue also does not divide interlocutors neatly into a speaker and an addressee and there are few occasions when participants produce a long monologue. Similarly, many utterances produced in a dialogue are only interpretable in the context in which they arise. Dialogue often exhibits what, in isolation, would be considered less than perfect examples of language. Many utterances in a conversation are not even grammatically well-formed sentences, yet people still manage to understand each other. There is therefore a necessity to investigate language in dialogue as well as in traditional psycholinguistic environments.

Research using dialogue paradigms has shown that language use is a highly interactive activity. Thus, addressees play a role in aiding the speaker to produce utterances, rather than alternating between production and passive comprehension (e.g. Bavelas, Coates and Johnson, 2000). Together interlocutors produce descriptions of objects or situations which are exclusive to a given interaction.

For example, Clark and Wilkes-Gibbs (1986) had one participant, the director, describe tangram figures (abstract geometric shapes) to another participant, the matcher, who had to put her tangrams into an order described by the director. Over the course of the conversation the director's descriptions of repeated tangram figures became shorter, more definite, and quite idiosyncratic. In addition, interlocutors tend to converge on similar descriptions, as we shall see.

2.2 Interactive Alignment Model

Pickering and Garrod (2004, see also Garrod and Pickering, 2004) have proposed the interactive alignment model of dialogue. Under this account alignment is the basis for successful communication, meaning that successful communication is characterized by interlocutors having sufficiently similar representations. Over the course of a conversation interlocutors align their linguistic representations, which leads to alignment of relevant aspects of the situation model, so that they are both using the same underlying representations to produce utterances. In this case, situation models are a multidimensional representation of the situation under discussion (Zwaan and Radvansky, 1998). The situation model is active in working memory and is a dynamic representation of information about the main characters under discussion, time, space, causality, and so on. Over time the situation models of interlocutors converge and the interlocutors become aligned. The speaker's situation model is primarily a representation of his or her own information, not a representation of what the addressee is likely to know or what is shared between the speaker and addressee. However, when the speaker and addressee are well aligned, it is also a good representation of the addressee's state of knowledge.

Alignment is achieved via three processes: (1) an automatic mechanism of alignment involving priming at all levels of linguistic representation and percolation between these levels; (2) a mechanism that repairs alignment failure; (3) alignment via explicit reasoning and modelling of a partner's mental state. This last process is used as a last resort when the automatic alignment processes fail. The interactive alignment model therefore acknowledges that interlocutors use low level automatic methods of aligning their representations (in process 1) as well as explicit modelling (in process 3), but the emphasis is on the use of automatic priming mechanisms to build up implicit common ground, which is the information that is shared between the interlocutors (see also Pickering & Garrod, 2006; Garrod & Pickering, 2004, 2007).

The interactive alignment model contrasts with theories of dialogue which argue that interlocutors model (full) common ground, which is the information that the interlocutors believe is mutually known (e.g., Clark 1996). On such accounts, the interlocutors regularly update common ground (thus carefully

demarcating what is mutually known and what is exclusively known). In this account a speaker introduces information which the addressee tacitly accepts either through saying things such as *yeah*, *ok* or by providing a new contribution. If the information is accepted then it enters common ground. If the addressee queries the information, then the two work together to resolve the issue, so that the original information or its replacement can enter common ground.

Clark's (1996) concept of common ground between interlocutors contrasts with Pickering and Garrod's (2004) concept of shared information (or implicit common ground). Common ground refers to a representation that each interlocutor has which contains the information that is available to both. It is explicitly marked as common ground and is explicitly maintained and updated as a conversation progresses. Common ground is a separate representation of knowledge independent of each interlocutor's own knowledge. Conversely, shared information, as suggested by Pickering and Garrod (2004), is not maintained and updated in its own right. Instead, shared information is simply the part of the situation model that overlaps between two (or more) interlocutors. As a conversation progresses the amount of shared information increases through the automatic priming mechanism.

2.3 Alignment of Representations

Pickering and Garrod (2004) argued that interlocutors largely achieve alignment of their situation models via alignment of linguistic representations and that this is achieved via an automatic priming mechanism. There is a great deal of evidence that interlocutors do mimic each other and tend to produce utterances that are similar to their partner's utterances and that this occurs at many levels of representation (e.g. lexical, syntactic, or semantic). Garrod and Anderson (1987) had pairs of participants play a cooperative maze game, in which they took turns to describe their positions to each other. The descriptions that the participants produced suggested that they were aligning situation models via the alignment of linguistic representations. Although there were several different ways in which participants could describe their position in the maze, for example a path description such as *I'm one along, and five up*, or a coordinate system such as *I'm at B5* (the maze was presented in a grid format), they tended to use the same description scheme as each other. That is, there was much more consistency within than between pairs in the method of location description. Participants also aligned on other aspects of their descriptions—for example, how they named different components of the maze. Some pairs referred to the locations in the maze (which were squares on the screen) as *nodes* whilst others referred to them as *boxes*. Once again there was greater consistency within than between pairs. The assumption is that the different ways of referring to the maze reflect differences in the underlying situation models

of the interlocutors and that, by aligning their language, pairs aligned the situation models that they were using to produce utterances about the maze (see also Garrod and Doherty, 1994).

Alignment is also evident in interlocutors' use of syntactic structures. A speaker's choice of syntactic structure has been argued to reflect an individual's situation model (Goldberg, 1995) and the evidence shows that interlocutors also align syntactic structures. Branigan, Pickering, and Cleland (2000) had participants describe pictures showing a dative event (the transfer of an object from an agent to a patient) and showed that a participant's use of a double object phrase (DO; e.g. *The waiter giving the customer the food*) or prepositional object phrase (PO; e.g. *The waiter giving the food to the customer*) to describe a target picture was influenced by the structure used by a scripted confederate to describe a prime picture: participants were more likely to use a DO construction after hearing the confederate use a DO construction than after hearing the confederate use a PO construction. This sort of alignment has also been shown for other types of syntactic structure, such as complex noun phrases (e.g. *The sheep that is red* vs. *The red sheep*; Cleland and Pickering, 2003) and has been shown when alignment leads to the production of syntactically ambiguous phrases (Haywood, Pickering, and Branigan, 2005). There is also evidence that bilingual interlocutors align syntax between L1 and L2, suggesting that syntax is shared between languages (Hartsuiker, Pickering, and Veltkamp, 2004; Schoonbaert, Hartsuiker, and Pickering, 2007).

Interlocutors have also been shown to align syntactically in spontaneous speech. Gries (2005) analysed a speech corpus and found that different dative constructions were more likely to be preceded by dative constructions of the same form, even when many contributions intervened between the two dative constructions. This shows that alignment does occur in natural speech, and also that the alignment effect persists over a relatively long time (see also Bock, Dell, Chang, and Onishi, 2007)

Evidence from syntactic alignment studies has also demonstrated another important feature of the interactive alignment model, namely that alignment percolates between levels of representation. This is shown by increased levels of syntactic alignment when there is lexical repetition across the prime and the target. Branigan *et al.* (2000) found that participants were more likely to use the same syntactic structure (DO or PO) if the verb used in the target was the same as the verb used by the confederate (e.g. *give* and *give* vs. *give* and *pass*). For example, participants were more likely to say *The waiter giving the customer the food* after hearing the confederate say *The nurse giving the patient the pill* than after *The nurse passing the patient the pill*. Similarly, Cleland and Pickering (2003) found that the semantic relatedness of the nouns in the prime and targets affected the level of alignment for complex noun phrases (e.g. *sheep* and *goat* vs. *sheep* and *car*). For example, participants were more likely to describe a picture as *The sheep that's red* after hearing the confederate say *The goat that's red* than after hearing *The car*

that's red. In both of these cases, lexical alignment enhanced syntactic alignment (see also Schoonbaert *et al.*, 2007).

Interlocutors also align lexically, in that they tend to use the same word as each other to refer to the same item. Brennan and Clark (1996) showed that interlocutors align or converge on the use of the same word to describe an object (e.g. *pennyloafer* vs. *docksider*). Garrod and Anderson (1987) showed that players in their maze game tended to describe their positions in mazes by using the same terms as each other, and giving those terms the same meaning (for example, *level* to mean *row*, counting from the bottom). In addition interlocutors also align on several other aspects including accent and speech rate (Giles and Powesland, 1975) and non-linguistic factors such as foot rubbing, face touching and body posture (Chartrand and Bargh, 1999; Shockley, Santana, and Fowler, 2003).

2.4 **Partner Specificity**

Brennan and Clark (1996) argued that interlocutors form conceptual pacts, whereby they tacitly agree to use a specific term to refer to a particular entity. It follows that the pact may exist only for that pair of interlocutors (or even only for that conversation). Therefore, when people switch conversation partners they have to form new conceptual pacts. In accord with this, Brennan and Clark found that speakers were more likely to retain a particular term when they subsequently interacted with the same partner than with a different partner.

However, the extent to which partner-specificity occurs is controversial, at least in comprehension. Barr and Keysar (2002) found that participants in an experiment were faster to look at objects the second time the object was mentioned, but that it did not matter who mentioned the object the second time (that is, whether it was a new or the old speaker); see also Kronmüller and Barr (2007). In contrast, Metzing and Brennan (2003) used a similar method that measured eye movements and did find partner-specific effects. Their experiment included a condition where a new term was used to refer to a previously mentioned object. Metzing and Brennan found that the addressee took longer to look at the object when an old speaker used a new term than when a new speaker used a new term. They argued that an old speaker's use of a new term to refer to a previously mentioned object broke a conceptual pact. The addressee, therefore, assumed that the new term referred to a new object, which caused processing difficulty. However, a new speaker's use of a new term to refer to a previously mentioned object did not break a conceptual pact, and so the addressee did not experience difficulty.

It therefore appears that partner-specificity does affect comprehension and production to some extent. According to the simplest automatic account, linguistic representations are activated by comprehension or production, and thus any subsequent act of comprehension or production is equally facilitated. However,

Garrod and Pickering (2007) point out that the interactive alignment model can account for partner-specificity by drawing on Horton and Gerrig's (2005) claim that people associate the use of terms with a particular speaker via implicit memory. If a speaker uses a particular expression, the expression and the speaker become associated in the addressee's memory. Therefore, the mention of that expression also activates information about the person who used that expression; and talking with the person who originally used that expression activates that expression in memory. Such a mechanism does not require conceptual pacts to explain partner-specificity.

2.5 Perspective Alignment

Producing or comprehending a spatial description requires an individual to adopt a perspective on a scene. For example, if two people are viewing a scene of a knife and a fork from opposite sides, then from one person's perspective the knife is left of the fork, whereas from the other person's perspective the knife is right of the fork. However, the speakers are not restricted to taking their own perspective: both can adopt the other person's perspective, and produce a description which corresponds to that person's perspective. For communication to be successful, it is important that the addressee adopts the same perspective as the speaker. Thus, we would expect perspective alignment to occur, as part of the general process of alignment of situation models (e.g. Garrod and Anderson, 1987).

In fact, people appear to readily adopt another person's perspective when describing the location of objects. Schober (1993) used a task in which a director and a matcher each viewed, from different perspectives, a large circle containing two smaller circles. On the director's scene one of the smaller circles was marked. The director had to describe which of the two was marked so that the matcher could mark that circle on his or her scene. Because the director and the matcher were viewing the scene from different positions the director could choose to describe the scene from his or her own perspective (i.e. egocentrically) or from the matcher's perspective (i.e. allocentrically). Schober found that directors were more likely to use an allocentric (i.e. matcher-centred) perspective than an egocentric one, suggesting that they were trying to minimize effort for their partner. In addition, because feedback from the matchers was allowed in the experiment it was possible to identify which perspective the matchers used when they questioned the director further about which circle to mark. The results indicated that the matchers used predominantly a director-centred perspective, which for the matcher was also an allocentric perspective. Therefore, when the directors spoke they tended to use the perspective of their addressee (i.e. the matcher) and when the matchers spoke they also tended to use the perspective of their addressee (i.e. the director). In this experiment the participants tended to align on using their addressee's perspective. However, this meant that the person whose perspective

was used switched depending upon whether the matcher or director was speaking. Directors used an allocentric perspective more when they were describing object location to a partner who was not present and so the director could not receive any feedback.

In Schober's (1995) study, the results showed that directors often used the perspective of the matcher when describing the location of the relevant objects. Interestingly, however, the directors did not use their partner's perspective exclusively, but occasionally used their own perspective and most often used neutral descriptions that were not from any perspective (e.g. *near*, *between*). Even though directors changed perspectives during the experiment, matchers were able to successfully complete the task without difficulty, thereby indicating that the matchers were able to understand directors despite the varying perspectives that the latter adopted. Clearly, then, matchers were able to understand which perspective the director was adopting for a given utterance, and to interpret the utterance accordingly. But how is it that an addressee is able to understand which perspective a speaker is using to describe the location of objects? Similarly, how is it that a speaker decides which perspective to use when describing the location of an object to an addressee?

2.5.1 Reference frame parameters

Given that alignment occurs at multiple levels of representation, Watson, Pickering, and Branigan (2004) hypothesized that interlocutors should also align spatial representations, and specifically the perspective that they use to describe the locations of objects in a scene. The perspective that is used to describe an object's location is dependent upon the reference frame that a speaker imposes upon the scene. A reference frame defines an origin, orientation, direction, and scale (Logan and Sadler, 1996). The origin refers to the object (or part of object) that the reference frame is situated on; this is usually the reference object (or part of it). For example, if Figure 2.1 is described as *The dot is above the chair*, the origin of the reference frame is on the chair. The orientation parameter determines which axes of the reference frame are the top-bottom, left-right, and front-back axes. For example, the above description of Figure 2.1 uses an intrinsic reference frame (defined below), and therefore the reference frame is oriented so the top-bottom axis is aligned along the canonically vertical axis of the chair. In this case this is the axis which extends from the chair legs through the seat and up to the top of the back of the chair. The direction parameter is nested within the orientation parameter; it sets the directional endpoints of each of the axes after the orientation has been determined. Continuing with the above example, once the top-bottom axis has been oriented in reference to the chair, the direction of the axis is assigned. In this case, the end of the axis that coincides with the canonical top of the chair is labelled as the top and the opposite end is labelled as the bottom. Finally, the scale parameter sets the distance according to one of the components of the scene.

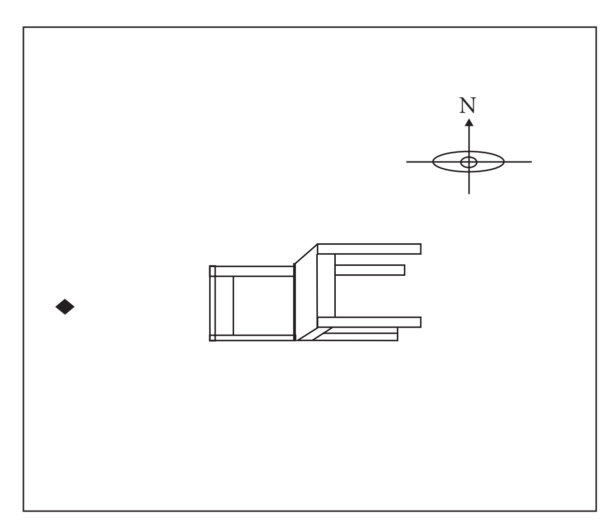


Fig. 2.1. The dot can be described as *west* of the chair using an absolute reference frame, *above* using an intrinsic reference frame, or *left* using a relative reference frame

The scale can be set according to the size of the chair (reference object), which is somewhat larger than the dot, or it can be set according to the size of the dot (figure object), which is the smaller of the two objects. Alternatively, both the figure and reference object may contribute to defining the scale parameter.

2.5.2 Reference frame types

The literature on spatial language has typically suggested there are three different types of reference frames: absolute, relative, and intrinsic (e.g. Levinson, 2003; Logan and Sadler, 1996). We illustrate these with respect to Figure 2.1.

Absolute reference frame. This refers to fixed features of the environment such as the points of the compass or gravity (or to directions such as downwind or inland in some cultures). In Figure 2.1, the dot can be described as *west of the chair*.

Intrinsic reference frame. The position of the figure in relation to the reference object is interpreted with respect to the actual orientation of the reference object. For example, the intrinsic meaning of *above* is (roughly) nearer to the top of the object than to any other part of it. As a chair has a top, the intrinsic reference frame allows the dot in Figure 2.1 to be described as *above the chair*.

Relative reference frame. The position of the figure object in relation to the reference object is interpreted with respect to the viewpoint of an observer. Using this reference frame, the dot in Figure 2.1 can be described as *to the left of the chair* (assuming the page is being held in a canonical fashion). In this case, the observer

is the reader. However, the relative reference frame can be used with different observers.

2.6 Investigation of Alignment of Reference Frames

In a series of experiments, Watson, Pickering, and Branigan (2004) have shown that interlocutors align reference frames when describing the location of objects to one another. Watson *et al.* used a confederate priming paradigm (Branigan *et al.*, 2000; Cleland and Pickering, 2003; Hartsuiker *et al.*, 2004) in which a confederate described the location of a dot (figure object) in relation to a reference object to a naïve participant, using either an intrinsic reference frame or a relative reference frame (prime). The naïve participant then chose which of two pictures matched the description given by the confederate (match phase). One picture, the *match scene*, matched the confederate's description according to either the intrinsic or relative reference frame; the other picture, the *distracter scene*, did not match the confederate's description according to either reference frame. The participant then described the location of a dot to the confederate (target), in the belief that the confederate had to choose which of two pictures matched their description. Several conditions for alignment have now been investigated in this way—of which two will be discussed here.

2.6.1 Within-axis alignment of reference frames

Experiment 1 investigated within-axis alignment of a reference frame between interlocutors (see also Carlson-Radvansky and Jiang, 1998). In this experiment the figure object was either in the same position on the target as the prime (same-preposition condition) or in the opposite position on the target as the prime (antonym-preposition condition). In the same-preposition condition, using the same reference frame as the confederate required the participant to use the same preposition as the confederate. For example, if the confederate described the match scene in Figure 2.2a as *The dot left of the chair* (using an intrinsic reference frame) and the participant used the same reference frame as the confederate, then the participant would describe the target as *The dot left of the chair.* In the antonym-preposition condition, using the same reference frame as the confederate required the participant to use the antonymous preposition. Therefore, if the confederate described the match scene in Figure 2.2b as *The dot right of the chair* (using an intrinsic reference frame) and the participant used the same reference frame as the confederate, then the participant would describe the target as *The dot left of the chair.* This situation was analogous for the other prepositions and the relative reference frame.

The results showed that participants used an intrinsic reference frame on 36.4% of trials after the confederate had used an intrinsic reference frame, but on only

Confederate's description and condition | Match and distracter scenes | Target scenes

a The dot below the chair—relative/absolute
The dot left of the chair—intrinsic
(same, vertical)

The dot below the chair—relative/absolute
The dot left of the chair—intrinsic

b The dot above the chair—relative/absolute
The dot right of the chair—intrinsic
(antonym, vertical)

The dot below the chair—relative/absolute
The dot left of the chair—intrinsic

Fig. 2.2. **a** represents the same-preposition condition; **b** represents the antonym-preposition condition (the match scenes are underlined solid, the distracter scenes are underlined dashed)

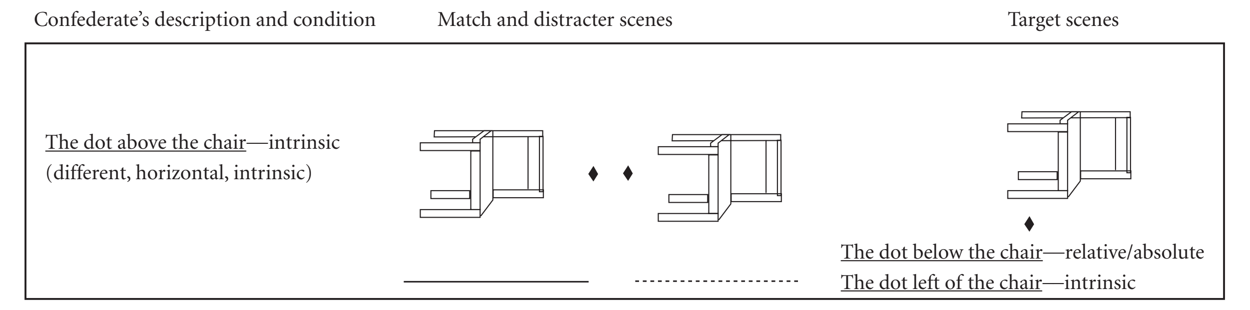


Fig. 2.3. The different-preposition condition for Experiment 2, which replaced the antonym-preposition condition in Experiment 1, shown in Figure 2.2b

26.9% of trials after the confederate had used a relative reference frame. This represents an alignment effect of approximately 10%. Hence participants were more likely to use a reference frame when it had just been used by the confederate than when the confederate had just used an alternative reference frame.

Experiment 1 only showed that interlocutors aligned single axes of reference frames (within-axis alignment). Thus, if one interlocutor used *above* in a relative reference frame, then a second interlocutor was more likely to use *above* in a relative reference frame or *below* in a relative reference frame; because *above* and *below* are on the same axis (vertical), alignment was within axes. It did not show whether alignment occurs between axes. Specifically, if one interlocutor uses *above* in a relative reference frame, would a second interlocutor be more likely to use *left* or *right* in a relative reference frame? Because *left* and *right* are located on a different axis (horizontal) to *above* and *below* (vertical), alignment in such a case would show that it occurs between axes.

2.6.2 Between-axis alignment of reference frames

Experiment 2 investigated whether participants would show between-axis alignment of reference frames, in addition to within-axis alignment of reference frames. In this experiment the confederate used intrinsic left and right in the same way as the participants in the previous experiment. One condition was identical to the same-preposition condition in Experiment 1, such that the figure object was in the same location on the prime and target (shown in Figure 2.2a). In another condition (different-preposition), the figure object was situated on different axes on the match and target scenes. Hence, if the figure object was located on a horizontal axis on the match scene, it was located on a vertical axis on the target, and vice versa. For example, if the confederate described the match in Figure 2.3 as *The dot above the chair* (using an intrinsic reference frame) and the participant used the same reference frame as the confederate, then the participant would describe the target as *The dot left of the chair*. If participants were more likely to use a reference frame after hearing the confederate use a reference frame in the different-preposition condition it would be strong evidence that interlocutors align the entire reference frame and not just the axes of a reference frame. This would also rule out lexical priming as an explanation of the alignment effect.

The results showed that participants used an intrinsic reference frame on 45.3% of trials after the confederate used an intrinsic reference frame, but on only 34.1% of trials after the confederate used a relative reference frame. As in Experiment 1, participants were more likely to use a reference frame if it had just been used by the confederate; furthermore, the effect was of a similar magnitude to Experiment 1. Hence, participants tended to align an entire reference frame. Importantly, in both experiments it was possible to establish that this was a true reference-frame priming effect and not a lexical priming effect. In both experiments, on half of the

trials, using the same reference frame as the confederate required the participant to use the same preposition as the confederate, but, on the other half of the trials, using the same reference frame as the confederate required the participant to use a different preposition. If the observed effects were lexical priming effects, then we would expect to find alignment only on those trials that involved the use of the same preposition in the prime and the target; if lexical priming contributed to, but was not wholly responsible for, the observed effects, then we would expect to find greater alignment in those trials which involved use of the same preposition in the prime and the target compared to those that involved the use of different prepositions on the prime and the target. However, the alignment effect was of the same magnitude (approximately 10%) when the prime and target involved different prepositions as when the prime and target involved the same prepositions. Hence, there did not appear to be any lexical contribution to the observed alignment effects. This observation is in keeping with work on syntactic priming where repetition of prepositions did not enhance priming (Bock, 1989).

2.6.3 Discussion

These results demonstrate that speakers' choice of reference frame during production of an utterance is influenced by the reference frame that they adopted during comprehension of their partner's previous utterance. Hence participants align reference frames during dialogue, just as they align other representations that support linguistic communication; moreover, the results demonstrate that this alignment effect is transient, with speakers able to switch between reference frames during the course of a conversation in response to their interlocutor's most recent choice of reference frame. This may account for why the alignment effect is only 10%. In other studies (e.g. Branigan *et al.*, 2000) the alignment effect has been much larger. However, in the majority of these studies fillers were used between experimental items. In Watson *et al.* (2004) there were no fillers, and instead the reference frame used by the confederate was fully randomized. This might therefore mean that the target was affected by previous utterances as well as by the prime. If this is the case then it is likely that using filler trials between experimental trials will increase the alignment effect and that in real language use reference frame alignment is greater than demonstrated in Watson *et al.* (2004). This is also supported by evidence that there is intra-speaker alignment of reference frames in addition to inter-speaker alignment (Vorwerg, this volume). One interlocutor will use one reference frame first. The second will align with this reference frame through inter-speaker alignment; this will then work in partnership with intra-speaker alignment to maintain a reference frame use throughout an interaction. However, the fact that interlocutors can align on a trial-by-trial basis indicates that flexibility of reference frame use can be maintained when the situation demands (Filipi and Wales, this volume), for example if the figure and

reference object are in a prototypical relationship for use of a specific reference frame.

These results are in keeping with Pickering and Garrod's (2004) interactive alignment model of dialogue. Alignment of reference frames is strong evidence that interlocutors align their situation models. Reference frames are abstract representations and the effects demonstrated in the two experiments are not directly due to alignment of linguistic representations. Indeed in the experiments there was no evidence of a lexical boost for reference frame alignment. That is, repetition of the preposition across prime and target did not increase reference frame alignment. Aligning reference frames is beneficial for interlocutors because it helps them to avoid potential misunderstandings: if they tend to re-use (in production) the reference frame that they have just comprehended, and assume that the other person is likely to do the same, they are more likely to interpret each other's ambiguous utterances correctly.

2.7 **Spatial Language in Dialogue**

The results also demonstrate that spatial language can be investigated using dialogue paradigms. This allows the examination of the representation of reference frames in a relatively naturalistic context. Previous experiments have used primarily monologue paradigms to investigate spatial language. The reference frames used in such experiments have been highly constrained by the monologue context (e.g. Carlson-Radvansky and Irwin, 1993; Carlson-Radvansky and Jiang, 1998; Logan and Sadler, 1996). This prevents any flexibility in reference frame use occurring as a result of dialogic processes between interlocutors.

Investigating spatial language using dialogue methods allows an examination of the effects of comprehension on subsequent production. In the results of Watson *et al.* (2004) it was shown that comprehension of an utterance using one reference frame affects the choice of reference frame for subsequent production of locative utterances. Further, their results show that interlocutors are willing to use locative descriptions that might be considered unusual outside a specific interaction (e.g. intrinsic left and right in Watson *et al.*, 2004). Such routinization has been shown in other areas of language, in particular lexical choice where interlocutors use interaction-specific words to refer to ambiguous figures (Clark and Wilkes-Gibbs, 1986), and locations in complex scenes (Garrod and Anderson, 1987). Intrinsic left and right is used relatively rarely; however in Watson *et al.* (2004) it appears that, because the confederate used the intrinsic reference frame, participants could become entrained to parse certain regions of space around the reference object in certain ways, even for the difficult left-right axis. It is clear from this that dialogue is an important method for the investigation of spatial language and that more naturalistic settings may induce further unusual uses of reference frames

which will help refine our understanding of the representations underlying spatial language.

2.8 Conclusion

Most research on spatial language has used monologue settings in which the use of reference frames has been constrained by the context of the experiment (with the notable exception of Schober, 1993, 1995). Such scenarios are not ideal to observe the use of reference frames in normal conversation. The results of Watson *et al.* (2004) show that interlocutors affect each other's use of reference frames. Specifically, comprehension of a sentence using one reference frame increases the likelihood of production of an utterance using the same reference frame. This is in line with other dialogue research showing that interlocutors influence each other's utterances (Pickering and Garrod, 2004) and shows that reference frames are amenable to investigation using dialogue paradigms.

3

Spatial Dialogue between Partners with Mismatched Abilities

MICHAEL F. SCHOBER

3.1 Introduction

A fundamental fact of spatial dialogue is that speakers can have different perspectives on what they are talking about: what is on one speaker's right can be on the other's left, or in front or behind (in back in US English), for the other. Of course, this is part of a larger truth about dialogue: beyond having different views of what is in front of them (their heads can never occupy precisely the same space, unlike Steels and Loetzsch's robots, this volume), speakers can have different world views, different agendas, different conceptualizations of the topics at hand, and different beliefs about what their conversational partner believes (see, for example, Russell and Schober, 1999; Schober, 1998a, 2005). Speakers thus confront the philosopher's notorious 'problem of other minds' every time they talk with each other, whether they know it or not, at multiple levels. At the level of spatial language, every use of a locative expression reflects a choice of reference frame or coordinate system, and thus provides evidence about how the speaker has attempted to solve the other minds problem at the current moment.

To consider this in a concrete setting, imagine that two people are facing each other over dinner at a restaurant. Jeff arrived early and ordered two different kinds of white wine, reflecting their different tastes. Aimee arrived later to find two similar-looking filled glasses on the table. The glass that is on Jeff's right is on Aimee's left. If Aimee asks which glass has the Riesling in it, Jeff has a number of choices, each of which reflects a different solution to the other minds problem. Jeff could avoid the problem by pointing silently at one glass, or pointing and saying 'that one.' But if he would rather not point in public, he must rely on language. He can say it's 'the one on your right', which not only uses Aimee's reference frame but marks it as Aimee's with the term 'your'. He can say that it's 'the one on the right', which takes Aimee's perspective but—riskily—assumes much more about Aimee's mental state: he assumes that Aimee will know that it must be Aimee's right that is being referred to. He can say it's 'the one on the left,' which takes a different risk in assuming that Aimee will know that Jeff is using his own perspective, or 'the one on my left,' which explicitly marks whose perspective is being used.

Of course, Jeff has the insurance plan that comes with participating in a dialogue (and not monologue): if Riesling-seeking Aimee has any doubts about whether she has understood the description she can ask for clarification ('Which one?' or 'You mean *your* left?').

Or consider a slight variation: the dinner companions are seated at a corner booth, with a 90-degree discrepancy in their viewpoints on the scene. Depending on how the glasses are laid out on the table, the terms 'right' and 'left' might well work from both parties' perspectives (what I have called a 'both-centered' perspective, Schober, 1993), and so there might be less need to mark whose left or right is being described. Other arrangements of objects and dinner companions will afford different potential descriptions, and will bring along with them ways of marking how much is assumed about what the partner knows. If, for example, there were three glasses as potential referents, Jeff could say that it is the one 'in the middle,' which would be true no matter where he and Aimee were sitting; this sort of 'neutral' perspective (Schober, 1995) can allow speakers to avoid having to choose one or the other's perspective. If any objects on the dinner table have their own coordinate system (front, back, top, bottom, left, right)—say, an unusual breadbasket, or a model plane—Jeff could use an object-centred perspective to refer to the desired glass as behind or to the left of that object.

Now, we already know that a number of factors affect spatial perspective choice. As I just noted, different physical settings afford different perspectives—objects' relative locations and the nature of the objects themselves (whether or not they have their own coordinate systems) allow different kinds of spatial descriptions (e.g. Levelt, 1982; Ullmer-Ehrich, 1982; Wunderlich, 1981). The degree of offset of the partner's vantage point can affect a speaker's perspective choice (e.g. Bürkle, Nirmaier, and Herrmann, 1986; Herrmann, 1989; Herrmann, Bürkle, and Nirmaier, 1987). The speaker's social goals (politeness, respect, egalitarianism) can affect perspective choice: college students are more likely to take a professor's spatial perspective than a fellow college student's (e.g. Graf, described in Herrmann, 1989), and people are likely to take their partner's perspective just as much as their partner has taken theirs (Schober, 1993). In earlier work (Schober, 1993, 1995, 1998b), I have shown that conversational feedback can affect perspective choice—people are more likely to take the perspective of a (silent) imaginary partner than of a live partner who can give feedback of understanding. And people are likely to stick with perspectives already taken with their partner.

Here I argue that there are additional differences between dialogue partners that can affect spatial perspective choice. The more I have thought about the issues, the more negotiating spatial perspective in dialogue strikes me as hard to divorce from negotiating the rest of what speakers negotiate in dialogue: their agendas (small talk or serious talk?) and thus their need for precision, their likelihood of requesting clarification, their desired level of politeness, and their assessments of each other's relevant abilities. Speakers have been observed to accommodate to each other on multiple dimensions, and there are many unanswered questions

about how this accommodation works and how individually variable it is (see Schober and Brennan, 2003, for discussion). Here I will focus on one of these factors—mismatched spatial abilities—and describe some results from a study (Schober, 1998c) that demonstrates how ability mismatches can drive perspective choice and affect the dialogue more generally.

3.2 **Study**

Why should spatial abilities affect linguistic perspective choice? At a fundamental level, taking one's partner's spatial perspective linguistically requires being *able* to see (or at least describe, which may or may not be the same) things from that point of view. How easy or hard this is to do turns out to vary across individuals (and along multiple dimensions which are correlated—see Hegarty and Waller, 2004, among others). If Aimee finds it extremely easy to imagine what Jeff is seeing, then it should be particularly easy for her to use language that reflects Jeff's coordinate system when describing a location for Jeff, and she should find it particularly easy to understand a description from Jeff's point of view. If Jeff has a terrible time imagining what Aimee is seeing, then it should be particularly difficult for him to produce and to comprehend a spatial description from Aimee's point of view. When Aimee and Jeff meet up, their relative abilities are, I propose, likely to affect what happens in the dialogue.

This hypothesis is lent plausibility by earlier evidence (Graf, cited in Herrmann, 1989) that college students are more likely to take the perspective of an (imaginary) small child than that of a fellow college student, presumably because they are less confident that the child will be able to understand descriptions that don't take their point of view. It is consistent with the finding that when speakers lack evidence about whether their partners have understood their spatial descriptions, they are more likely to take the addressee's perspective (Mainwaring et al., 2003; Schober, 1993). The hypothesis is also lent plausibility by the many demonstrations of how speakers adapt to their less capable partners in other arenas. Normally abled caretakers adapt to their conversational partners with mental retardation on several dimensions (Abbeduto, Weissman, and Short-Meyerson, 1999; Testa, 2005). Adults with greater expertise tailor their utterances for their partners with lesser knowledge (Isaacs and Clark, 1987), and even children adapt their speech to younger children with less linguistic ability (Shatz and Gelman, 1973).

The experiment described here tests this hypothesis by asking people to describe locations for partners. Unbeknownst to them, participants had been selected for having very high or very low mental rotation abilities as assessed by performance on a timed mental rotation test, the Card Rotations Test (S-1 rev.) from the ETS Kit of Factor-Referenced Cognitive Tests (1976/1992). Of course, the abilities that a mental rotation test taps are disputed; it is not clear whether test-takers mentally rotate the figures or themselves—see, for example, Just and

Carpenter, 1985; Wraga, Creem, and Profitt, 2000; Zacks, Mires, Tversky, and Hazeltine 2000—or whether they carry out imagistic or propositional or some other sort of transformation—see, for example, Pylyshyn, 2002. But the intuitive notion here was that whatever abilities a mental rotation test taps are likely to be necessary for spatial dialogue when partners have different points of view on a scene.

Participant selection. Participants were not informed about how they had been selected for the study; the preselection test was part of a packet of questionnaires given long before the study to over 700 students in large Introductory Psychology classes at SUNY Stony Brook. The test consisted of two sets of 80 items which ask whether two simple figures at different rotations are the same or different (reversed); students were given three minutes to attempt each set of 80. Scores on the test are typically normally distributed. For the purposes of this study, students who averaged above 54 out of 80 on the two tests were considered to be of high ability, and students who scored below 37 out of 80 were considered to be of low ability.

Experimental task. Based on these scores, 70 pairs of students were brought to the laboratory. In each pair, one student, randomly selected to be the *director*, was seated in front of a computer monitor; his or her task was to describe locations on a series of 32 displays (presented in different random orders for different pairs), which were designed so that each locative description could unambiguously be coded as reflecting only one perspective (round 1). The other student, the *matcher*, marked locations in a paper packet with the same 32 displays. After doing this, the students switched roles and described 32 more displays (round 2).

The displays were designed to require participants to take one or another perspective from among five possible perspectives as described in Schober (1995): speaker-centred, addressee-centred, both-centred, object-centred, and neutral. Speaker-centred descriptions are true from the speaker's point of view and not from the addressee's, and addressee-centred descriptions are true from the addressee's point of view and not the speaker's (whether or not they are explicitly marked for speaker, as in 'my left' or 'your left'). Both-centred descriptions are true from both the speaker's and the addressee's point of view (e.g., 'on the left' when both parties are looking from more or less the same angle) and are not otherwise marked for which speaker's perspective is intended; only some arrangements of speakers allow both-centred descriptions, but facing one another does not. Object-centred descriptions (or 'intrinsic' descriptions; see Watson *et al.*, this volume) reflect the coordinate system or metaphorical point of view of a non-human object that has its own front and back or left and right, as automobiles and other vehicles do; of course, such descriptions may reflect the imagined point of view of a human using the vehicle, but the critical point is that they reflect the object's orientation and perspective independent of the human observers (that is, no matter where speakers are in relation to the car or each other, the exhaust pipe is still at the back end of the car). And neutral descriptions are other descriptions

that are independent of the human observers' points of view but that do not reflect the intrinsic perspectives of the objects in the scene; for example, if a target object is 'between' two cars, this will be the case no matter which direction the cars or the observing humans are facing.

As in Schober (1993), the displays were minimal and simple, consisting of objects without their own coordinate system (circles) in various arrangements, sometimes including an additional simple object with its own clear front, back, left, and right (an aeroplane). The idea was to create a set of stimuli (1) with the minimum complexity necessary for distinguishing between these different perspectives; (2) that contrasted addressees and objects in four different rotations (0, 90, 180, 270 degrees of disparity); (3) that alternated which circle was the target object, and (4) that allowed for unambiguous descriptions that could clearly be coded for one perspective only. Obviously, these scenes substantially underestimate the complexity and ambiguity of possible scene arrangements in the real world, but they represent a reasonable set for experimental purposes.

The final set of 32 displays (sampled from the 61 logically possible displays from our combination of elements) included eight two-circle displays that required partners to use either the speaker's or addressee's perspective (see Figure 3.1a); eight that allowed speaker-centred, addressee-centred, and neutral description (see Figure 3.1b); eight that allowed speaker-centred, addressee-centred, and object-centred descriptions (with the addition of the outline of an aeroplane) (see Figure 3.1c); and eight that allowed not only speaker- and addressee-centred but also neutral and object-centred descriptions (see Figure 3.1d).

To exemplify, in Figure 3.1a speakers must refer to the target circle either as the one on the right or the left, which reflects a choice of one party's perspective ('right' reflects the speaker's perspective and 'left' reflects the addressee's); displays in this series were arranged so that there were never indeterminate views that might allow both-centred descriptions. In Figure 3.1b, the speaker could refer to the target circle as the one in front of the others (speaker's perspective), behind (in back of) the others (addressee's perspective), or the one in the middle (neutral perspective). In Figure 3.1c, the target circle is on the left (speaker-centred), on the right (addressee-centred), or behind the plane (object-centred description). And in Figure 3.1d the target circle is in the left row of circles, the second one back, from the speaker's perspective; in the back or far row, second from the left or right, from the addressee's; on the right from the plane's (object-centred) perspective; and in the middle circle in the row of three or next to the wing (true from any perspective, and thus neutral). To locate the target circle in this more complex scene, the speaker could easily use more than one locative description, potentially reflecting more than one perspective, as in 'it's the middle circle (neutral) on my left (speaker-centred)'. In this scene a speaker could even use a both-centred perspective if they were to refer to the bottom-left circle as the circle 'on the left', which is true from both the speaker's and addressee's point of view, and indeterminate as

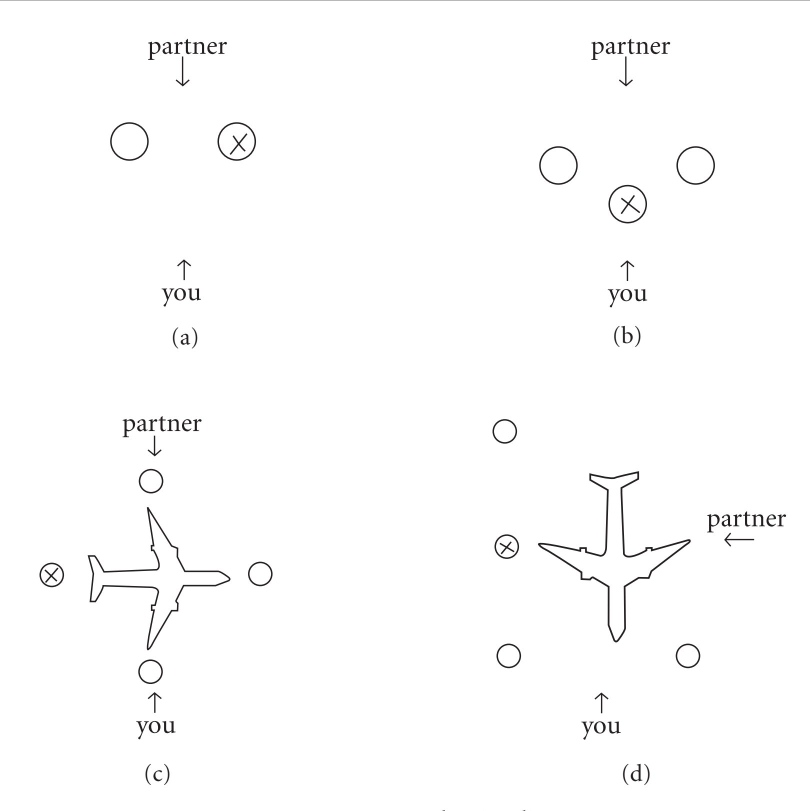


Fig. 3.1. Sample stimuli

to which it is. But the target circles were selected so as to avoid such indeterminacies. So in some of the displays there were as few as two possible perspectives for participants to use (director's vs. matcher's), and in some there were as many as five.

After the experiment, each participant again took the timed mental rotation test to verify their ability level, because low scores in preselection could reflect laziness or inattention rather than actual ability. After eliminating the surprisingly high number of pairs whose scores were inconsistent with the earlier test, the conversations of the 29 remaining pairs were transcribed and coded for perspective. In this final sample, the mean score of high ability participants was 72.1, ranging from 54.5 to 80, and the mean score of the low ability participants was 19.2, ranging from −3.5 to 36.5 (a negative score meant that the participant went slowly and got more comparisons wrong than right).

Transcription and coding. Audiotapes of the conversations were transcribed. Each locative description was coded by two coders (the few disagreements were resolved by discussion) for whether it reflected the director's perspective, the

matcher's perspective, a both-centred perspective (true from both interlocutors' points of view and not marked as belonging to one or the other), an object-centred (the plane's) perspective, or a neutral perspective (true from anyone's point of view). Any descriptions that did not reflect any interlocutor's (or any discernible) point of view were coded as bizarre.

Note that the taxonomy of perspectives reflected in this coding scheme differs from the more frequently employed distinction between intrinsic and relative perspectives, which in most cases treats all observer-centred perspectives as equal rather than distinguishing between speaker- and addressee-centred perspectives (although see Tenbrink, 2007, for schemes that make finer-grained distinctions). The taxonomy is also task-specific, in that the stimuli in this experiment—both the target objects to be described and the objects as arranged—were selected so that there would never be a possible descriptive scheme for target objects that was ambiguous with respect to these perspectives. In real-world scenes it is regularly possible for speakers and objects to be situated in ways that allow descriptions that are ambiguous with respect to perspective; if, for example, the plane in Figure 3.1c were facing one of the partners and the target circle were in front of or behind the plane, the description 'behind the plane' would be ambiguous with respect to whether it reflected the plane's object-centred (intrinsic) back or one of the interlocutors' points of view. With the arrangement in Figure 3.1c, the description 'behind the plane' unambiguously reflects an object-centred frame of reference (which holds for both speakers), while a description of 'on the (my) left' or 'on the (your) right' would reflect one of the interlocutors' points of view. Similarly, for Figure 3.1d, a description of the target circle as 'to the left of the plane' would reflect the speaker's perspective (as seen from 'you'), while 'behind the plane' would have to reflect the addressee's perspective, and not an object-centred (plane's coordinates) perspective, because the target circle is not at the plane's own back end. So it is the particular arrangements of interlocutors, display objects, and circles selected as targets that allowed for the unambiguous coding. (Of course, this also affects the generalizability of the results, in that it is unknown how frequently real-world settings allow interlocutors, or researchers, to set perspectives so unambiguously).

Note also the importance of the displays being presented within an experimental task that made comprehension unambiguously measurable, and independent of any coding or judgement of the suitability of locative expressions (see, for example, Carlson and Hill, this volume; Tenbrink, 2007, and this volume, for discussion of the communicative principles that guide object reference in contrastive scenarios). The criterion for comprehension accuracy was whether the matcher marked the appropriate circle on the display, whether or not the director's description might be considered ambiguous by an outside observer; in many other domains interlocutors come up with schemes for understanding each other that are not transparent to outsiders (see, for example, Schober and Clark, 1989), and I would argue that there is an important distinction to be made between codability

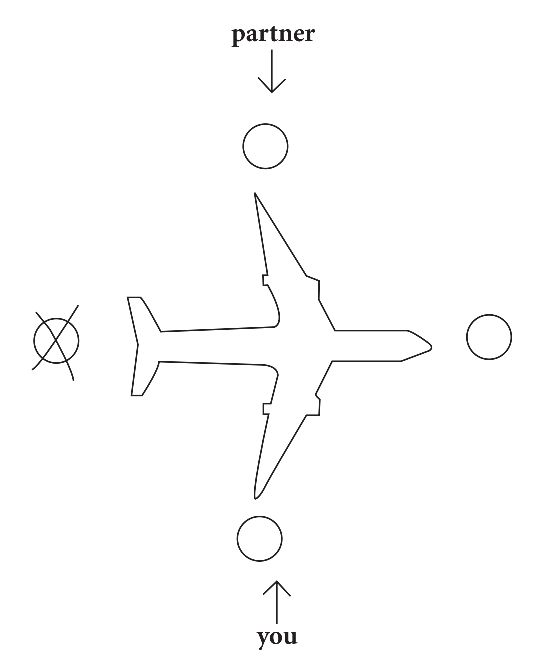


Fig. 3.2. Matcher's point of view on display, Round 2 #1

of a description as an overhearing outsider and comprehension accuracy for an addressee.

Results. What was immediately apparent in the transcripts was how much the partners' relative mental rotation abilities affected the nature of the interaction, despite the fact that participants were entirely unaware that they had been selected on the basis of their relative abilities. Compare the following exchanges, which were the first interchanges in the second round of interaction, after the interlocutors already had experience with the other in the alternate task role; in both cases the matcher saw the display in Figure 3.2.

The first is a typically brief interchange between a high-ability director paired with a high-ability matcher:[1]

D: It's behind the plane (OBJECT-CENTRED)
M: Okay

The matcher selected the correct target circle, despite the fact that she could have interpreted this as one of the other circles; clearly this pair had come to agreement about the use of object-centred perspectives during their earlier interactions.

[1] In the transcript excerpts, D refers to the Director and M to the Matcher. Overlapping speech is enclosed in asterisks and pauses are indicated by periods between spaces. Coded perspectives are indicated in capitals.

The second exchange, describing the same display, is a far more tortuous interchange between a high-ability director and a low-ability matcher:

D: Okay my plane is pointing towards the left (DIRECTOR-CENTRED)
M: Okay
D: And the dot is directly at the tail of it (NEUTRAL)
D: Like right at the back of it (OBJECT-CENTERED)
M: Okay mine is pointing to the right (MATCHER-CENTRED)
D: Oh yours is pointing to the right (MATCHER-CENTRED)
M: Yeah
D: So your dot should be on the left (MATCHER-CENTRED)
D: Because my dot is on the right (DIRECTOR-CENTRED)
D: In back of it (OBJECT-CENTRED)
D: So your dot should be at the left (MATCHER-CENTRED)
D: At the back of it right (OBJECT-CENTRED)
M: Yeah
D: Yeah
M: But if it is the same—but if it—the same dot-right? Wait a minute, if my—your plane is pointing to the left *[something]—*
D: *My* plane is pointing to the left (DIRECTOR-CENTRED)
M: Mm-hm
M: And that dot and the dot that's highlighted is the one all the way in the back of it (OBJECT-CENTRED)
M: Like behind the tail (OBJECT-CENTRED)
M: Yes, so so my dot is gonna be
D: So my dot is on the right (DIRECTOR-CENTRED)
D: And yours should be on the left right (MATCHER-CENTRED)
M: Yeah
D: Okay *so your—*
M: *Right behind the tail* okay (OBJECT-CENTRED)
D: Okay

Simple counts of the numbers of words that each pair needed to accomplish their task demonstrate that directors in high-high pairs used marginally fewer words than directors in either mixed-ability or low-low pairs, $F(1,26) = 3.60$, $p < .07$ (see Figure 3.3).

Another striking finding was that low-ability interlocutors were far more likely to produce bizarre uncodable descriptions like this one, produced by a low-low ability pair for the display in Figure 3.4:

D: Um the top circle (BIZARRE)
M: There is no top
D: Oh wait um the bottom circle (DIRECTOR-CENTRED)
M: There's no bottom

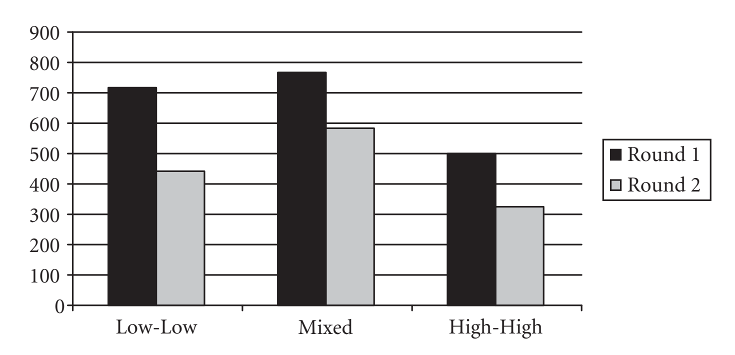


FIG. 3.3. Mean number of words spoken by directors by group

D: Oh the circle to the left left (BIZARRE)
M: Okay

Note that the director seems to be stabbing in the dark, hoping for a lucky description that the matcher will find acceptable; note also that the low-ability matcher accepts a description that simply isn't true from either party's perspective and ends up marking the wrong circle. As Figure 3.5 shows, low-low pairs produced more bizarre descriptions than mixed pairs, and mixed pairs produced more than high-high pairs, linear trend $F(1,26) = 5.74$, $p < .03$.

This stabbing in the dark pattern, and the seeming inability to decentre from their own points of view, was evident throughout the low-ability pairs' conversations, as in this example (see Figure 3.6):

D: The top circle (BOTH-CENTRED) to the right (DIRECTOR-CENTRED)
M: There's no top circle at the right
D: Uh, oh the bottom the bottom circle (BIZARRE) to the left of of (MATCHER-CENTRED) the picture

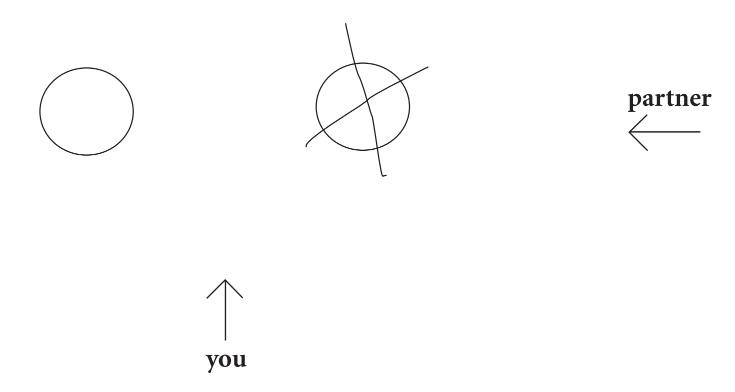


FIG. 3.4. Matcher's point of view on display, Round 1 #14

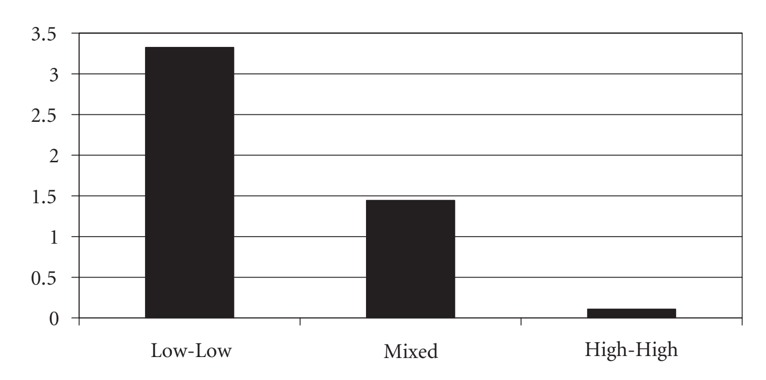


FIG. 3.5. Mean number of bizarre descriptions by group, Round 1

M: Um there's none at the left
D: Oh um the bottom circle (BIZARRE) to the right (DIRECTOR-CENTRED)
M: There's none at the right
D: Uh top circle (BOTH-CENTRED) to the left (MATCHER-CENTRED)
M: Okay

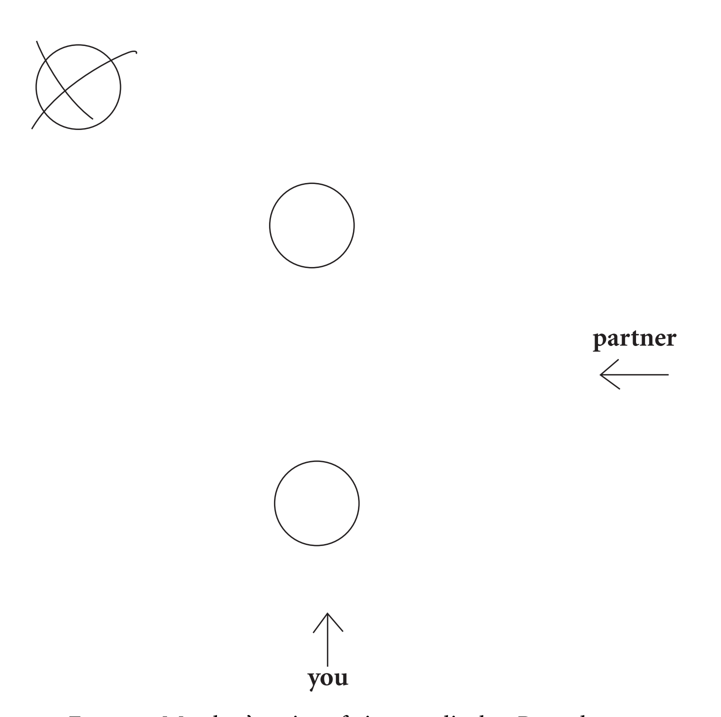


FIG. 3.6. Matcher's point of view on display, Round 1 #29

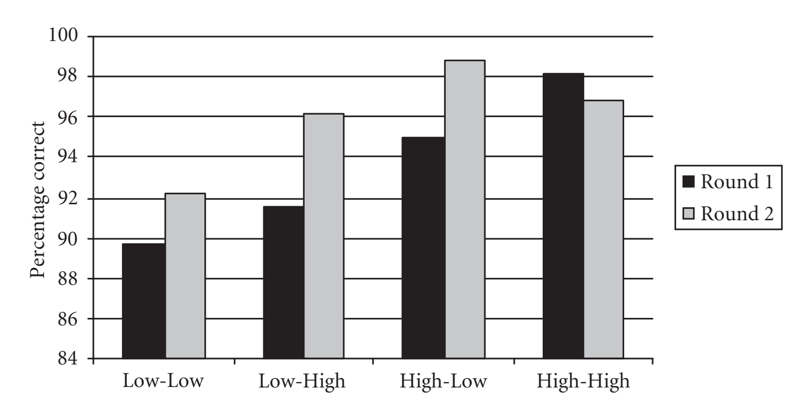


Fig. 3.7. Matchers' accuracy: mean percentage of correct circles marked across trials by group

Not surprisingly, matchers in low-ability pairs were rather inaccurate in marking the appropriate objects, indicating their generally poorer comprehension (see Figure 3.7). But the only reliable difference was between low-low and low-high pairs, $F(1,25) = 4.47$, $p < .05$; the comprehension for matchers in mixed pairs seemed to be protected by having one high-ability partner—whether matcher or director.

What about taking the other person's perspective? The very clear finding is that low-ability directors were more likely to take their own perspective, while high-ability directors were more likely to take the matcher's perspective, interaction $F(1,25) = 9.31$, $p < .005$ (see Figure 3.8). If we examine matcher-centred descriptions in further detail, we see that their use changed over the course of the 32 descriptions. As Figure 3.9 shows, if we compare the first eight and the final eight

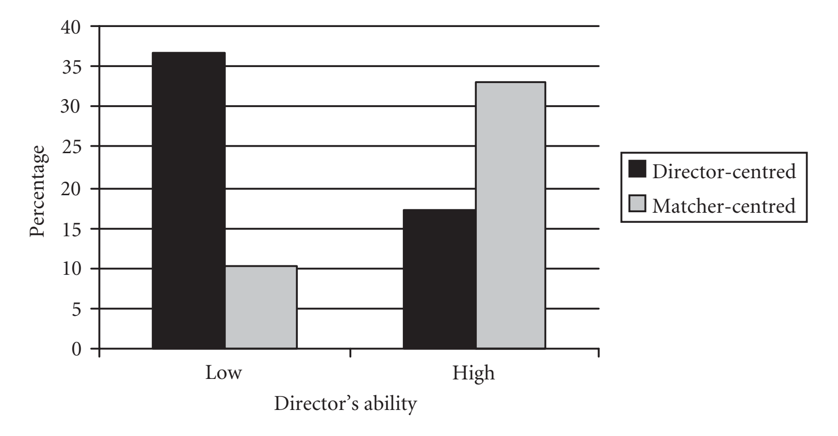


Fig. 3.8. Perspectives, Round 1

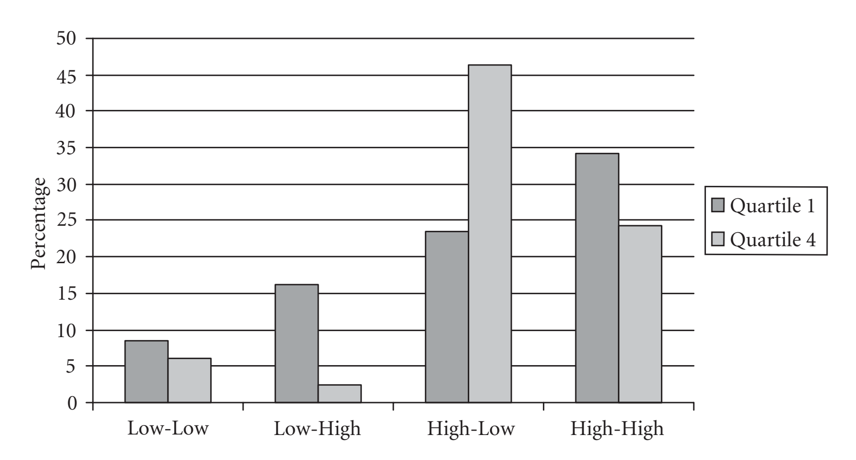


FIG. 3.9. Matcher-centred descriptions, Round 1

descriptions, we see notable changes in the use of matcher-centred descriptions, interaction $F(1,25) = 11.21$, $p < .005$, as if partners were getting to know each other's abilities as the round went along. Already in the first quartile, high-ability directors were more likely to take their partner's perspective. But as time wore on in mixed pairs, high-ability directors with low-ability matchers drastically increased their use of matcher-centred perspectives. Low ability directors with high-ability partners notably decreased their attempts at matcher-centred perspectives.

Closer examination of the transcripts suggests part of what is going on: high-ability directors recognize that their partners are having trouble understanding director-centred descriptions and switch to taking their partner's point of view more and more. And low-ability directors end up being licensed by their high-ability matchers' questions to use more director-centred descriptions, as in this example describing the display in Figure 3.10:

D: Okay now, it's uh the second one (BOTH-CENTRED)
M: The one towards the bottom of the page? (MATCHER-CENTRED)
D: I'm sorry?
M: The one towards the bottom of the page (MATCHER-CENTRED)
M: Or the top? (MATCHER-CENTRED)
D: Uh there are two circles near
M: I oh yours are right next to each other (DIRECTOR-CENTRED)
D: Yeah
M: All right the one towards
D: It's
M: The left (DIRECTOR-CENTRED)
M: Or right? (DIRECTOR-CENTRED)
D: Left (BIZARRE)

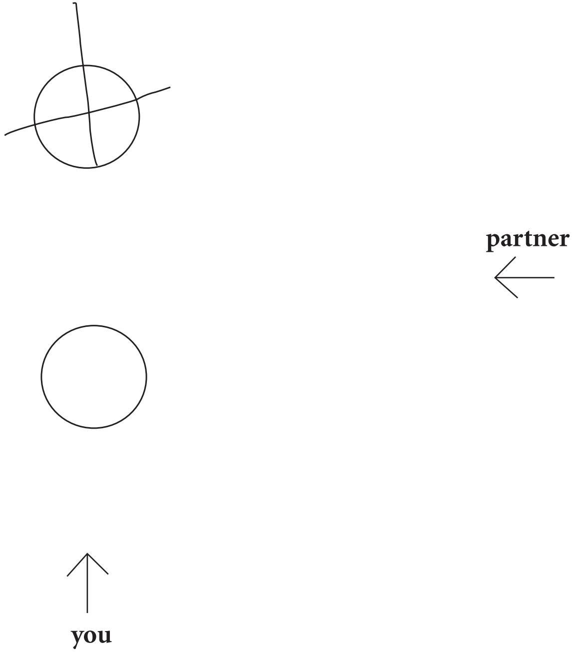


Fig. 3.10. Matcher's point of view on display, Round 1 #21

M: All right *got it*
D: *Okay* mm-hm
M: Yeah

Directors' and matchers' abilities did not affect their use of neutral and object-centred descriptions; the only findings with these perspectives were that directors of all abilities were more likely to use neutral and more object-centred descriptions at 90- and 270-degree offsets than when they were face to face, replicating the general pattern in Schober (1995) with a different set of displays.

Overall, what emerges from these findings is not only the extent to which the shape of the conversation is affected by individual abilities but also the extent to which the conversational context—the partner's uptake and input—shapes what each individual does. As has often been argued elsewhere, language in interaction is simultaneously an individual-mind and collective phenomenon, with social processes affecting individual cognition and individual cognition affecting social processes. Either level of analysis without the other leaves out something important.

More particularly, the findings demonstrate that not only individuals' spatial abilities but also their interlocutors' abilities can substantially affect people's choice of perspectives in describing locations, and that they subtly judge

each other's abilities within a few moments of conversing, by as yet unknown mechanisms. Pairs in which both partners had poor mental rotation abilities understood each other's spatial language more poorly than pairs in which at least one partner had high mental rotation ability, which seemed to allow pairs to compensate for the difficulties that people with low ability had. People with low ability were far more likely to provide ineffective or uncodable descriptions that were not true from anyone's perspective. Speakers with high ability were more likely to take the perspective of partners with low abilities, essentially encouraging them to speak egocentrically, and this propensity increased over time, as they gained further evidence of their partner's ineffectiveness.

3.3 Further Questions

Adding individual spatial abilities into theories of spatial language and dialogue raises a host of unanswered questions, just as the addition of individually variable cognitive capacities (like working memory span) has complicated theories of language processing more generally (see e.g. Kurtz and Schober, 2001; Miyake, 2001, among many others). First, these preliminary findings suggest that perspective choice in spatial language use is closely related to other aspects of spatial cognition—but which aspects and when? Given that different abilities, like spatial orientation and spatial visualization, are at least partially dissociable (Hegarty and Waller, 2004), how might the different abilities affect spatial language choice—and which aspects? At least under certain circumstances—certain spatial scenes, certain relative locations of partners—the use of spatial language by definition requires assessing the partner's point of view, and this requires some form of mental rotation. Might different rotation strategies that speakers take affect their spatial language use? The current findings only begin to suggest that the primary site of any effects will be in person-centred (speaker and addressee) perspectives, but this would obviously need to be studied with a greater range of scenes and in more naturalistic settings.

Second, much is unknown about how partners assess each other's abilities. The data here hint that a partner's failure to produce and comprehend sensible descriptions tips off high-ability speakers, but how does this work? Are particular kinds of failures particularly informative? Beyond varying in their abilities to produce and comprehend spatial descriptions, do speakers also vary in sensitivity to social cues, or empathy, or interest in perspective taking, in ways that interact with their spatial language choices? Ability assessments in other domains appear to be relatively quick, just as they were here: consider Isaacs and Clarks' (1987) New Yorkers who very quickly changed their level of detail about New York landmarks when they figured out their partners were non-New Yorkers. And there is some evidence that people can use the *way* their partners talk—the extent of pausing and disfluency, and doubtful tone of voice—to judge how confident their partners

are about what they are saying or how likely their partners are to need clarification (e.g. Brennan and Williams, 1995; Schober and Bloom, 2004). But is the process of assessing each other's relative abilities the same in spatial language as in other settings? And do people accommodate in the same ways?

Third, the current study selected participants whose abilities were on the extreme ends, and the differences in ability in mixed pairs were extreme. How do abilities come into play for less extreme abilities or for less extreme ability differentials? Note also that, because this study's participants' abilities were extreme, they would be likely to perform just as well or as poorly at the various correlated components of perspective taking. But to the extent that components are dissociable, how might those different components affect spatial language use?

Fourth, the findings here are necessarily limited in generalizability by the choice of stimuli and experimental setting. The stimulus displays were selected to allow unambiguous coding of perspective and clear measurement of the accuracy of description, and so the arrangements of scenes, stationary interlocutors, and target objects were carefully chosen, with recognition that they were odd. The experimental setting had the additional oddness of both parties' not seeing each other's faces, and looking at different representations of the same scene (one on a monitor, one on a piece of paper). Exactly how this all affected results is unknowable, and I would argue for future research that sets up situations that are closer to real-world settings. But this leads to the question of what real-world settings are really like and how they vary for different people and different realms of experience; as far as I am aware, it is entirely unknown how often speakers describe location in settings with so clearly definable perspective choices, how often the interlocutors remain stationary and can see each other's point of view, how often they use language (as opposed to pointing) to refer to target objects, and how often their descriptions allow their partners unambiguous access to whose perspective is being chosen.

Despite these limitations of the current study, I propose that spatial language provides an opportunity for researchers in other sorts of language and dialogue, because (in the right circumstances) the terms that interlocutors use allow clear assessment of who has taken whose perspective, what references speakers have intended, and what addressees have understood; in addition, researchers can manipulate factors that affect perspective-taking. This is an unusual opportunity for researchers interested in the problem of other minds who have focused on other sorts of dialogue; in other settings quite often one cannot tell whether people using the same words really have the same representations underlying them (Schober, 2005, 2006). That is, when two people are discussing 'abortion' or 'euthanasia' or even less loaded topics like their 'jobs', the mere fact that people use the same words is not a guarantee that they have the same underlying representation; lexical alignment can actually mask undetected conceptual misalignment, which can lead to miscomprehension and task failure (see, for example, Conrad and Schober, 2000; Schober, Conrad, and Fricker, 2004, for evidence from survey

interviews). The facts of spatial perspective actually provide a clear case where sometimes one person's use of 'left' means something quite different from the other's. This allows for empirical testing of claims about the extent to which two parties' conceptual alignment is automatic when the same words are used (Pickering and Garrod, 2004; Vorwerg, this volume; Watson *et al.*, this volume). As I have argued elsewhere (Schober, 1995, 1998a), it also allows for empirical testing of claims about the extent to which partners allocate effort within a pair as opposed to the individual, as in Clark and Wilkes-Gibbs' (1986; see also Clark, 1996) proposals about least collaborative effort, in a situation where each party's individual effort is definable.

Acknowledgements

This work was supported by NSF Grants SBR-97-30140, IRI-94-02167, and SES-05-51294, as well as a New School Faculty Development Award. I am grateful to Susan Brennan and the SUNY Stony Brook Psychology Department; to Jon Bloom, Silvia Leon, Marcy Russo, Catherine DiNardo, Danielle Barry, Yassi Gidfar, Jack Shelley-Tremblay, and Susan Berrill for their help in recruiting and running participants and coding the data; and to the anonymous reviewers and the editors for their useful comments.

4

Consistency in Successive Spatial Utterances

CONSTANZE VORWERG

4.1 Introduction

Several factors that influence the choice of a spatial reference frame have been identified in previous research. These include functional relations (Carlson, 2000; Carlson-Radvansky and Radvansky, 1996; Coventry and Garrod, 2004), parameters of the communication situation (Herrmann and Schweizer, 1998; Schober, 1993), the interlocutors' spatial abilities (Schober, this volume), perceptual saliency (Vorwerg and Kronhardt, 2008) and linguistic factors (Vorwerg and Weiß, 2008). However, people often produce more than a single verbal localization. In successive spatial utterances, the choice of a spatial reference frame—regardless of other factors—might be influenced by previous utterances. Indeed, an intra-individually consistent use of reference frames has been observed in a number of studies on connected discourse or localization sequences (Ehrich, 1985; Ehrich and Koster, 1983; Levelt, 1982; Vorwerg, 2001). It is not yet clear from these results whether the consistency found is really an effect of the earlier utterances on the later ones. First, one frame of reference—usually a deictic one—is predominant in many of these studies. Its predominance has been attributed to its suitability to be used in a consistent way—contrary to an intrinsic frame requiring reference objects to have intrinsic fronts (Ehrich, 1985). A consistent use of one predominant spatial perspective might also be observed if speakers resorted to default frames of reference (see Watson, Pickering, and Branigan, 2004, for a discussion regarding dialogue effects). Second, even when intra-individual uniformity of reference frame selection is found to be coupled with inter-individual variability, this pattern of results—as argued by Levelt (1982)—might be explained by (possibly in part genetically determined) cognitive styles underlying the use of spatial perspective.

Contrary to that notion, *consistency effects* can be put down simply to the influence of previous utterances provided that the reference frame selection can be shown to depend on initial items, by manipulating starting conditions. In this account, an inter-individually varying, but intra-individually consistent employment of spatial reference frames can be produced by a combination of (1) different

starting conditions and (2) a tendency to use frames of reference consistently in a given situation. Such a cognitive principle of consistency has been suggested to be effective in perceptual judgements, such as LARGE or HEAVY, requiring a scaling or categorizing frame of reference (Haubensak, 1992; for a discussion of the relation between cognitive consistency and social attitude see Eiser, 1971). Given that spatial reference frames used in spatial language are a special case of the broader notion of a frame of reference in perception and categorization (Vorwerg and Rickheit, 1998), it can be expected that such a principle of consistency holds for spatial frames of reference, too. Experimental evidence supporting this idea (Vorwerg and Kronhardt, 2008) is discussed in Section 4.2.

This review of consistency effects in reference frame selection provides the background for a presentation and discussion of new findings concerning some other aspects of spatial utterances, which are addressed in Sections 4.3 and 4.4. These findings from a free-verbal-localization experiment suggest that intra-individual consistency leading to striking differences between participants—obviously due to differences in the initial utterances—can also be observed for syntactic and lexical representations, which also tend to repeat across verbal-localization trials. First, Section 4.3 focuses on a range of linguistic means that are available to express a particular spatial relation verbally. A spatial concept, such as BEHIND, can be encoded by a preposition + noun phrase (e.g. *behind the box*), a preposition + pro-form (e.g. *behind it*), or an adverb (e.g. *behind*). In German, these different uses of 'behind' correspond to three different lexical items (*hinter—dahinter—hinten*) representing different word categories (preposition—prepositional adverb—adverb; for other syntactic forms of direction terms in English and German, see Tenbrink, this volume). The findings in Section 4.3 support the idea that individual consistency and the initially presented object locations play a role in choosing one of these spatial word categories.

Second, Section 4.4 focuses on another linguistic aspect in verbal localization, namely word order within complex spatial expressions. Often—for directions in between prototypical axes—a combination of two spatial relations (e.g. LEFT + BEHIND) is used to denote a location in space. To examine what factors contribute to the order in which the two possible dimensions (sagittal: front/behind, and lateral: left/right) are verbalized, the utterances were related to the spatial locations presented, classified according to syntactic relatedness, and compared between subjects. The results revealed that syntactically closely related dimension pairs (with one spatial term modifying the other) differed in their order between participants, while in other cases the order of spatial relations verbalized depended on spatial characteristics of the location named. Thus, Section 4.4 highlights the role of individual consistency for dimension order.

Both the consistency in reference frame choice and the newly observed consistency effects in lexical selection and the naming order of spatial dimensions probably reflect fundamental processes in dialogue, which may either be related to a conscious endeavour to be consistent in one's judgements and utterances,

or to an automatic activation of underlying mental representations at different linguistic levels. According to the interactive-alignment account advanced by Pickering and Garrod (2004), interlocutors adapt to each other in dialogue due to alignment through priming of those representations which are activated when comprehending an encountered utterance (see also Watson, Pickering, and Branigan, this volume). In the same way, the representations underlying one's own utterances might prime their later use and lead to a kind of intra-personal alignment in discourse. These questions are discussed in Section 4.5.

4.2 Reference Frame Choice

Vorwerg and Kronhardt (2008) examined in an experimental study the question of whether later utterances within a localization sequence are influenced by preceding ones with respect to reference frame selection. In this study, participants were asked to describe the location of a small wooden ring from a toy construction kit relative to a toy aeroplane constructed from the same kit (see Figure 4.1). In this situation, there are two possible spatial frames which can be used to localize the intended object. One possibility is to use the plane's intrinsic orientedness for determining the main axes constituting the frame of reference. In this case, an object located at the front of the plane would be described as *in front of it*. The other option is to use one's own spatial orientation towards the plane for determining the reference axes. In that case, an object located between the speaker and the toy plane would be described as *in front of it*.

The first variant is an example of a binary localization, in which the reference object's (relatum's) intrinsic orientedness defines the point of view from where a spatial relation is judged. Here, only two entities are needed to establish the spatial relation: the located object and the reference object. The alternative is an example of a ternary localization, in which a third object or entity other than the relatum is used as the point of view from where a spatial relation is judged (see Herrmann, 1990). The distinction between binary and ternary localization corresponds to Levinson's differentiation between 'intrinsic' and 'relative' relations (Levinson, 1996; see note 32).

However, a second criterion may be equally important for the classification of reference frames, namely the type of perspective used (see Herrmann, 1990), as this might be related to underlying perceptual or cognitive processing (Vorwerg, 2001; Vorwerg and Kronhardt, 2008). While our first localization possibility uses a perceived object's orientation to define the reference directions, our second localization variant employs a communicator's (here: the speaker's) perspective. For the purposes of the present chapter, a binary reference frame on the basis of a perceived object's orientation—as in the first case described—is called *intrinsic*, and a ternary reference frame constructed on the basis of the speaker's orientation—as in the second case described—is called *deictic* (see Table 4.1). In German, English, and many other European languages, the deictic reference frame is usually based

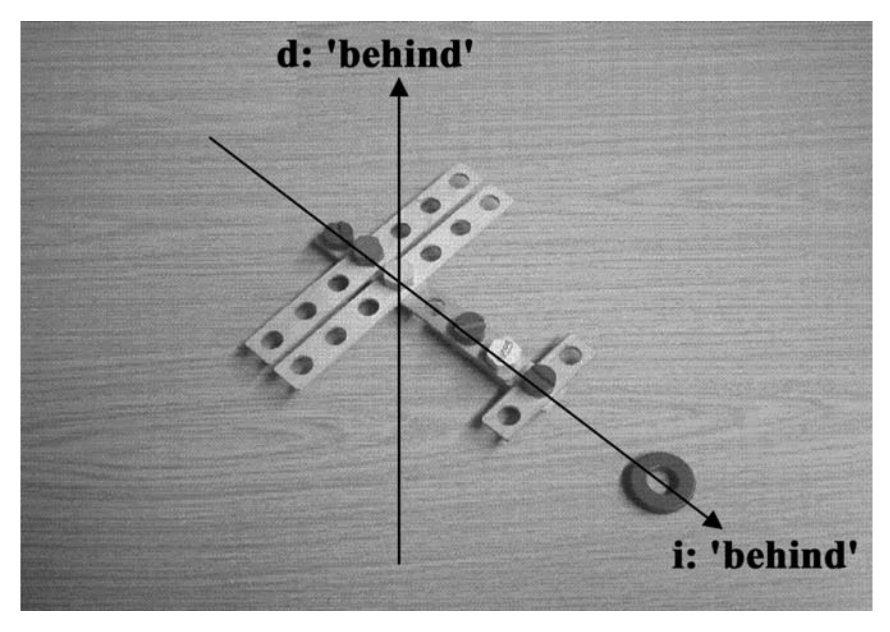


Fig. 4.1. Example item for the 'intrinsic' condition. Approximate deictic (d) and intrinsic (i) reference axes are indicated.

Table 4.1. *Taxonomy of frames of reference for direction specification based on Herrmann (1990). Assumed underlying perceptual frames are given in parentheses.*

Point of view	Binary localization	Ternary localization
Speaker (observer)	*egocentric* (body-centred)	*deictic* (viewer-centred)
Other entity	*intrinsic* (object-centred)	*extrinsic* (environment-centred)

Source: Vorwerg and Kronhardt, 2008.

on a "mirror" principle. That is, the orientation of the reference frame is "mirroring" the front, back, left, and right poles of the speaker's orientation in space, such that both are facing each other. Accordingly, the opposite principle—usually used in an intrinsic frame—may be called a "tandem" principle, as the orientation of the reference frame and of the reference object are aligned in this case.

In order to be able to investigate whether initial uses of the intrinsic as opposed to the deictic frame of reference would influence subsequent localizations, it was important to influence the initial choice by varying starting conditions; otherwise observed consistency effects could be due to 'default settings' or individual preferences. Vorwerg and Kronhardt (2008) manipulated the perceptual saliency of spatial positions within one or the other frame of reference during the first trials. We chose this kind of manipulation because it enabled us to address another issue

we were interested in: the question of whether perceptual saliency of a location in terms of direct locatedness at a reference axis (or prototypicality for a frame of reference) would enhance the probability of choosing this frame. To this end, we needed a design where the reference axes of both frames differ maximally. Therefore, the aeroplane presented as reference object was rotated by 45° relative to the participant's line of sight, such that proximal deictic and intrinsic axes always differed by 45° (see Figure 4.1).

The minimal spatial configurations consisting of the toy plane and the small ring were presented on a computer screen. Two different orientations of the plane were used, both differing from the line of sight defining the deictic frame by 45°. In one orientation, the plane was facing away (as depicted in Figure 4.1: ↖); in the other one, it was facing the speaker (rotated by 180° relative to Figure 4.1: ↘). The position of the ring differed between trials covering locations at all eight half-axes (four intrinsic and four deictic) plus all eight bisector directions between proximal axes. So, 16 locations around the reference object spaced at 22.5° were presented for each orientation of the plane.

A total of 96 native speakers of German participated in this experiment. Each subject produced two sequences of 16 localizing utterances each, one for the orientation of the plane shown in Figure 4.1 (↖) and one for the reversed orientation (↘). They were asked to name the location of the ring relative to the plane with a short spatial expression (such as *davor* 'in front of it') for each configuration presented. The crucial manipulation was the initial location of the ring at an intrinsic vs. a deictic reference direction (cf. Figure 4.1) in each sequence.

Results showed an effect of perceptual saliency as expected. In their initial localization, participants were more likely to use an intrinsic frame of reference if the located object was located at an intrinsic reference direction given by the main axes of the reference object. Otherwise, a deictic frame of reference was preferred. These data suggest that speakers are more likely to use that frame of reference on whose ideal-typical axis the intended object is located. (Other experimental results presented by Carlson and Hill, this volume, indicate that a prototypical relation with an intended object may also enhance the likelihood of an object being selected as a reference object if there is a choice of possible relata.)

This initial-selection effect provided the basis for investigating consistency effects. Only those items were included in the data analysis which could unequivocally be assigned to one of the two frames of reference in question. Results revealed a high consistency rate defined as the relative frequency of the particular dominant (i.e. the more frequent) frame of reference in a sequence (with a value range between 0.5 and 1.0). Mean consistency was .93 for a very conservative analysis with pre-defined categories based on the assumption of a mirror principle in deictic localizations and a tandem principle in intrinsic localizations. However, this value even underestimated the mean degree of consistency since some subjects used unexpected types of reference frames, such as left-right or front-back reversals (relative to the usual mirror and tandem

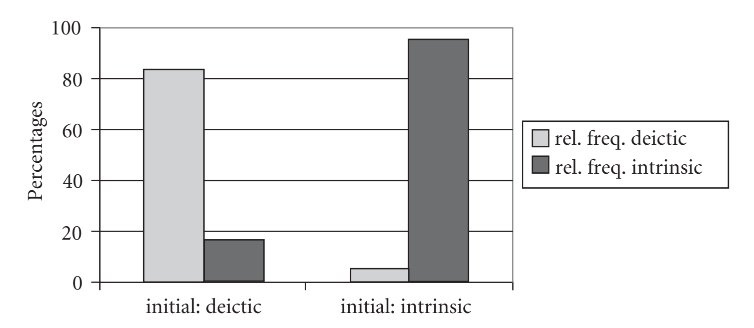


FIG. 4.2. Mean relative frequency of reference frame in sequences depending on initial use.

principles in deictic and intrinsic frames respectively), or extended categories. Taking into account these unusual frames of reference, the mean consistency was .96 (5% *trimmed mean* = .98). In 76% of localization sequences, one single reference frame was chosen over all trials.

Having established a consistency effect, the next question to ask was whether 'intrinsic consistency' or 'deictic consistency' (the consistent use of an intrinsic or a deictic frame of reference respectively) depended on the initial choice. Results revealed a significant effect of initial selection on subsequent items, as indicated by a strong correlation (*Spearman's* ρ = .84). The relative frequency of deictic or intrinsic localizations within a sequence depended very clearly on the initial choice of a frame of reference (see Figure 4.2). The slightly smaller value for initial deictic selections resulted from the fact that a small number of subjects changed to an intrinsic frame after the first or second item, as revealed by transition probabilities between successive localizations throughout the sequence. The number of predominantly deictic vs. intrinsic patterns also depended on initial locatedness. This confirms the conclusion that the initial locatedness on an intrinsic vs. a deictic axis influences the use of a particular frame of reference in later localizations.

Altogether these data reflect the tendency to adopt a similar frame of reference throughout a localization sequence. They suggest that the selection of a frame of reference is a function of previous spatial utterances within a sequence. Spatial representations activated during language production tend to repeat within a certain context. The cognitive mechanism leading to consistency in reference frame selection might either be a more or less conscious endeavour to be consistent across different trials by categorizing similar locations similarly or using spatial terms as before, or a simple priming mechanism by which previous representations pre-activate a frame of reference for the following localizations. This question will be taken up in the discussion section.

In any case, the fact that the use of a particular spatial frame of reference—which is to a large extent consistent across localizations within a sequence—depends on an externally manipulated factor suggests that patterns of spatial perspectives varying between language users can be a combined result of differing starting conditions and a tendency towards consistency in the use of reference frames.

4.3 **Lexical Choice (Word Category)**

A direction category to which a spatial relation can be assigned using a reference frame, such as BEHIND, can be linguistically encoded in a number of ways. While the linguistic form used to express a spatial relation is specified by design or instruction in many experiments, free language production data are needed to study factors of word category choice. Data from one such study have been analysed here in order to examine possible consistency effects in the lexical selection for localization. The complex data set had previously been analysed regarding factors of spatial categorization, abstracting away from linguistic means (Vorwerg and Rickheit, 1999). Based on a summarizing overview of the spatial language used (Vorwerg and Rickheit, 2000), I have reanalysed the data to look into the factors influencing word-category choice, specifically with regard to possible consistency effects. In this section, the choice of lexical category will be addressed; results concerning the order of combined direction terms are then reported in the section following.

Participants were presented with configurations of a die and a wooden slat from a toy construction kit. For each configuration, participants were asked to describe the location of a single die with respect to a slat. The object pairs were presented on a computer screen in 3D space using a pair of stereo glasses. In this 3D picture, both objects were lying on the same table plane. The slat was centred on the table plane and could have one of four possible orientations, two collateral to the deictic frame (either sagittally: |, or laterally oriented: —), and two rotated by 45° relative to this frame (either / or \, as seen from above). In each trial, the die could have one of 72 possible locations around the reference object, yielded by a combination of three distances and 24 directions. The 24 directions were defined relative to the orientation of the slat and covered the four half-axes, the eight extended edges in the horizontal plane, the four diagonals stretching from the corners, and the eight bisectors of the axis extensions and the diagonals (see Figure 4.4).

A total of 35 native speakers of German participated in the experiment, 18 of whom gave verbal descriptions with respect to the two collateral orientations, and 17 of whom gave verbal descriptions with respect to the two diagonal orientations of the slat. Altogether, each participant produced spatial utterances for 144 locations (24 angular positions $\times$ 3 distances $\times$ 2 orientations of the slat).

TABLE 4.2. *Overview of German directional ('projective') prepositions, adverbs, and prepositional adverbs. Additionally,* oberhalb *and* unterhalb *can be used as vertical prepositions or adverbs.*

Spatial dimension	Adverb	Preposition	Prepositional adverb
Vertical	*oben, unten*	*über, unter* [+ dative]	*darüber, darunter*
Sagittal (primary horizontal)	*vorn, hinten*	*vor, hinter* [+ dative]	*davor, dahinter*
Lateral(secondary horizontal)	*links, rechts*	*links (von/neben + dat.), rechts (von/neben + dat.)* [+ genitive]	*links davon, rechts davon, (links/rechts)* daneben

The order of items was randomized. Subjects were asked for *short* descriptions; otherwise, their answers were not restricted in any way.

The main linguistic means used to express *direction* were prepositions (in 22% of utterances), adverbs (60%), and prepositional adverbs (24%). Adjectives were used in 2% of utterances, usually related to a part (edge, end, side) of the relatum and—with few exceptions—in addition to another direction term. Some utterances (7%) included terms to denote *distance*; these were employed almost exclusively to supplement direction specifications.

The analysis of the factors influencing word-category selection was confined to (directional) prepositions, adverbs, and prepositional adverbs, as these cover almost all utterances (99.7%). In German, spatial terms belonging to one of these three word classes differ by word form (with exception of LEFT/RIGHT terms, which are either part of complex expressions or marked by case as prepositions). So, the preposition used for a BEHIND relation is *hinter*; the adverb used for the same relation is *hinten*. The prepositional adverb *dahinter* combines *da* ['there'] with the prepositional form *hinter* ['behind'] and is used as an adverb (meaning 'behind it'). An overview of prepositions, adverbs and prepositional adverbs for directional relations is given in Table 4.2.

The analysis of direction term distribution revealed that individual consistency is a main factor of lexical choice. Subjects tended to stick to the same wording pattern throughout the experiment. The absolute number of utterances per subject containing prepositions, adverbs, prepositional adverbs or some combination of them is given in Figure 4.3. Adverbs were counted as such if they were used to denote a spatial relation by themselves (e.g. *rechts hinten*), but not if they were used to modify a spatial term of another word category (e.g. *rechts dahinter*). However, in some utterances, several (simple or complex) spatial expressions were strung together (e.g. *hinten* < – > *links dahinter*).

Examples of the wording of individual participants are given in Table 4.3. As these examples illustrate, the localizing sequences of individual subjects can

Table 4.3. *Individual examples of participants' wording patterns*

Participant 18: Prepositions 'to the left of . . . '	Participant 33: Adverbs 'on the left'	Participant 3: Prepositional adverbs 'to the left of it'
links neben der Leiste	*vorne, links*	*rechts davor*
links hinter der Leiste	*links*	*davor*
links hinter der Leiste	*ähm links*	*links davor*
links hinter der Leiste	*links*	*links davor*
hinter der Leiste	*hinten, links*	*links davor*
hinter der Leiste	*hinten*	*links daneben*
hinter der Leiste	*hinten, Mitte*	*links daneben*
rechts hinter der Leiste	*hinten*	*links daneben*
rechts neben der Leiste	*hinten, rechts*	*links daneben*
rechts neben der Leiste	*hinten, rechts*	*links daneben*
rechts neben der Leiste	*hinten, rechts*	*links dahinter*
rechts neben der Leiste	*hm hinten, rechts*	*links dahinter*
rechts neben der Leiste	*hinten, rechts*	*links dahinter*
rechts vor der Leiste	*rechts*	*dahinter*
rechts vor der Leiste	*rechts*	*rechts dahinter*
vor der Leiste	*rechts*	*rechts dahinter*
vor der Leiste	*vorne, rechts*	*rechts daneben*
vor der Leiste	*vorne*	*rechts daneben*
vor der Leiste	*hm vorne*	*rechts daneben*

usually be distinguished by the patterns of language means used to encode directional relations. Patterns may also be characterized by the use of (weakening or strengthening) linguistic hedges (such as 'slightly', 'rather', or 'exactly')—employed to express grades of spatial-category membership (see Franklin, Henkel, and Zangas, 1995; Vorwerg and Rickheit, 1999)—or by typically embedding spatial expressions in whole sentences (such as, 'Now it is . . . '). Furthermore, speakers' utterances differed according to whether they employed specific subcategories of prepositions, adverbs, or prepositional adverbs. Examples are ABOVE/BELOW vs. IN-FRONT/BEHIND expressions for the sagittal axis (for a discussion of 2D conceptualization of 3D space with objects in one plane, see Vorwerg, 2001; Vorwerg and Rickheit, 2000), or the use of *oberhalb/unterhalb* vs. *über/unter* (or *oben/unten*). Another example is the use of 'dynamic' adverbs: *nach links leicht nach oben* ['to the left slightly up'] (depending on deviation from axes).

What can account for the observed inter-individual differences in lexical choice? Do they reflect individual styles or preferences? In order to find out whether lexical-choice patterns are influenced by starting conditions, data were aggregated for statistical analysis. Both the individual wording patterns and the fact that prepositional adverbs are a kind of pro-form substituting explicitly for a

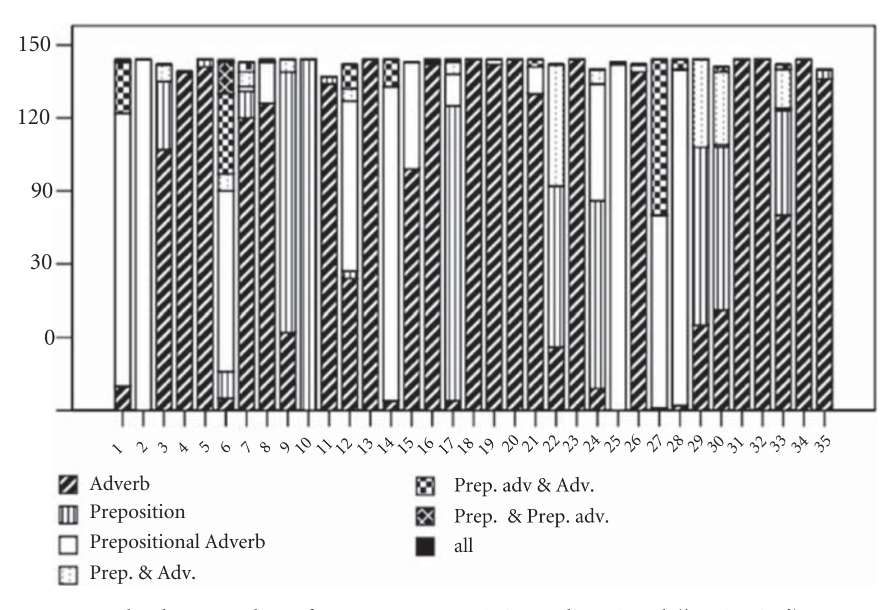


Fig. 4.3. Absolute number of utterances containing a directional ('projective') preposition, adverb or prepositional adverb or a combination of these, per subject.

prepositional phrase suggested that participants might differ fundamentally in whether or not they employ prepositions and prepositional adverbs. Therefore, utterances were classified according to whether or not they included one of these.

In order to address the question of whether there is a relationship between proportion of use of a word category and the directional relation between objects, it was necessary to differentiate between those participants who received two slat orientations collateral with the line of sight (n = 18) and those who saw two slat orientations rotated by 45° relative to the line of sight (n = 17). Only data referring to the collateral orientations (see Figure 4.4) were included in this analysis. The 24 locations used around the reference object represent six different types of deviation from axis and edge (see Figure 4.4).

As Figure 4.5 shows, the overall use of prepositions and prepositional adverbs depended on the combined directional proximity to the nearest axis and edge. In addition to individual consistency, deviation from reference lines (axes and extensions from edges) was one of the factors affecting the probability of a lexical choice.

In view of the fact that the order of items was randomized for each participant, the finding that the type of location affected the probability of choosing a 'prepositional expression' (preposition or prepositional adverb) led to the hypothesis that the initial item influenced the wording pattern throughout the experiment. First, to establish an effect of initial location on the initial word choice, a Chi Square test was performed relating location (0 to 2 vs. 3 to 5) and

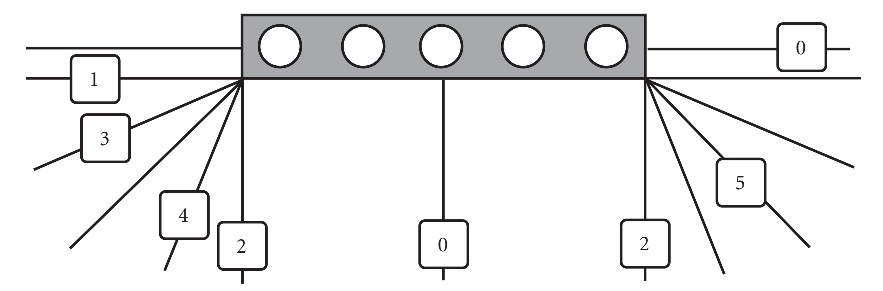


Fig. 4.4. Schematic depiction of locations around the reference object (as seen from above). Six types of deviation from axis and edge were used: (0) at an axis, (1) no deviation from edge with small deviation from axis, (2) no deviation from edge with large deviation from axis, (3) small deviation from edge with small deviation from axis, (4) small deviation from edge with large deviation from axis, and (5) maximal deviation (diagonal).

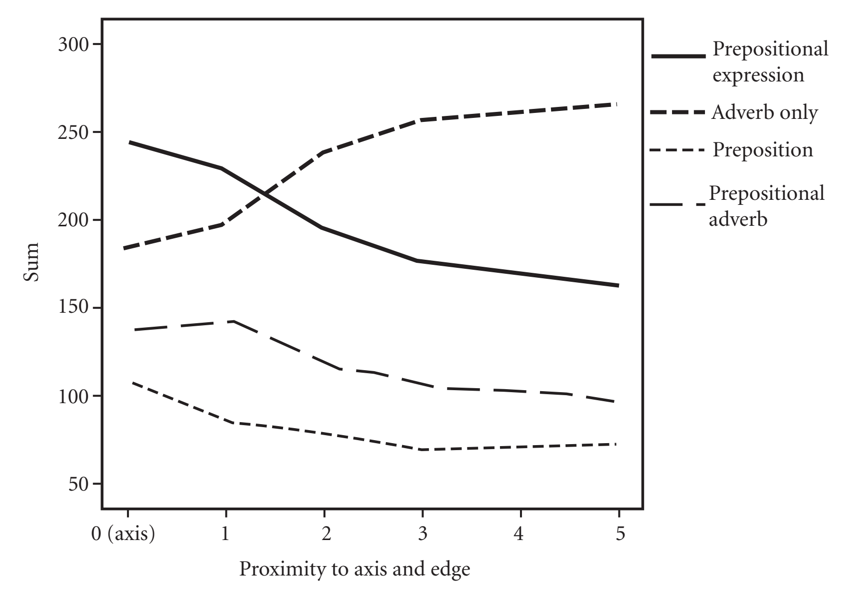


Fig. 4.5. Absolute number of utterances including a prepositional expression (preposition or prepositional adverb) or not (only adverbs) in relation to the type of location (defined by proximity to axis and edge, see Figure 4.4). Frequency of prepositions and prepositional adverbs contribute to 'prepositional expression'; shown here for purposes of comparison.

TABLE 4.4. *Relationship between first lexical choice and wording pattern (for sequences: n = 35; $\chi^2 = 16.3$; p < .01)*

	First utterance		
	prepositional	adverbial	*Total*
Pattern			
prepositional	8	2	10
adverbial	1	14	15
mixing	7	3	10
Total	16	19	35

lexical choice (with vs. without prepositional expression). It revealed a significant relationship (at the .05 level) between initial location and first utterance: out of seven participants starting with a proximal location (no. 0 to 2), six used a prepositional expression and one did not; out of eleven participants starting with a distant-from-axis location (no. 3 to 5), three used a prepositional expression and eight did not.

Second, to uncover whether the initial word choice affected later ones, individual localization sequences were assigned to three kinds of pattern: (1) predominantly prepositional, (2) predominantly adverbial, (3) mixing. A pattern was classified as predominantly prepositional if no more than 10% of utterances (14) were adverbial. A pattern was classified as predominantly adverbial if no more than 10% of utterances (14) were prepositional. Otherwise it was classified as mixing prepositional and adverbial utterances. Table 4.4 shows the relationship between first lexical choice and wording pattern (for all sequences [n = 35] in order to increase statistical power).

The statistical analysis revealed that the use of a prepositional or an adverbial pattern was affected by the initial utterance. Participants starting with a prepositional expression (preposition or prepositional adverb) tended to use it throughout the experiment, whereas participants starting with an adverbial description most often stuck to that. Therefore, it can be concluded that the observed relationship between the number of prepositional expressions and the combined proximity to axes and edges (see Figure 4.5) is mainly based on the data of those participants mixing perspectives (six out of the 18 subjects receiving collaterally-oriented slats). As Figure 4.6 shows, directional deviation seems to be a decisive factor in switching between prepositional and adverbial utterances, that is, diverging from a consistent pattern. This holds even for individual localization sequences and for word subcategories.

To sum up, there were different wording patterns for individual localization sequences. At the level of the broadest word category, prepositional, adverbial, and mixed patterns can be distinguished. The probability of using a prepositional

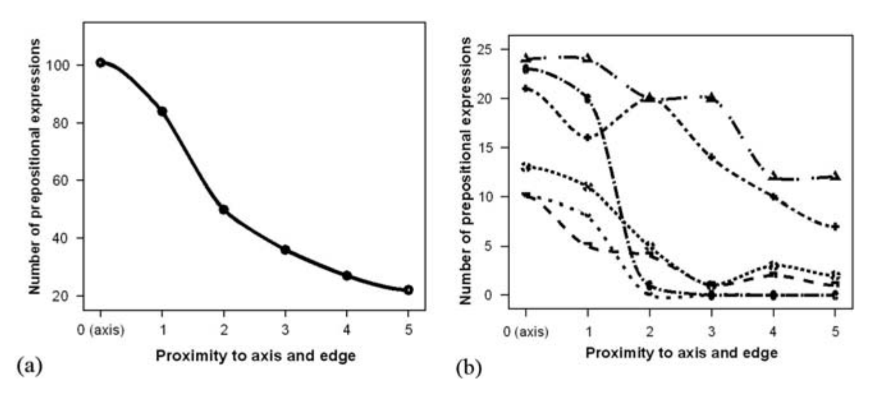


Fig. 4.6. (a) Relationship between frequency of use of a prepositional expression (preposition or prepositional adverb) and combined proximity to axis and edge based on data from those subjects using a mixed pattern. (b) Curves for individual participants.

vs. an adverbial pattern throughout the experiment was a function of the initial choice, which in turn was influenced by starting conditions (directional deviation). For those participants in the experiment mixing prepositional and adverbial utterances, diversions from a consistent pattern seemed to be related to directional deviation.

4.4 Syntactic Combination of Dimensions

In a sequence of verbal localizations related to the same kind of configuration, people often use combined or modified direction terms specifying both a lateral and a sagittal/vertical direction. One interesting question to ask is what principles govern the order of dimensions in those utterances. Sometimes one direction is specified primarily and another one is used to give additional details, such as in the formulation *hinten, etwas rechts* ('behind, a little to the right'). In these cases, the main directional category is produced first (a spatial-conceptual factor).

However, in German there is also a kind of 'syntactic combination' possible (e.g. *hinten links* 'behind-left' or *links hinten* 'left-behind'), in which the first term seems to modify the second. For these combined expressions, order of dimensions was analysed using the same data set as for the word-choice analysis presented in the previous section. Surprisingly, this analysis revealed that intra-individual consistency was also an important factor in determining order of dimensions (lateral—vertical/sagittal vs. vertical/sagittal—lateral). Eleven out of 35 participants in the experiment used the same order throughout (see examples given in Table 4.5), and an additional ten subjects used the same order with few exceptions (1–4). Altogether, most participants (33 out of 35) preferred one order

Table 4.5. *Individual examples of participants each using only a single order of dimensions*

Subject 4: vertical—lateral	Subject 35: lateral—vertical
unten links	*links unten*
oben rechts	*rechts unten*
unten links	*links unten*
oben rechts	*links oben*
oben rechts	*rechts unten*
oben links	*links unten*
oben links	*links oben*
unten links	*links unten, äh nein mittig*
oben rechts	*rechts oben*
oben links	*links oben*
unten rechts	*rechts oben*
oben links	*links oben*
unten links	*rechts unten*
oben links	*links unten*
unten li/unten rechts	*links oben*

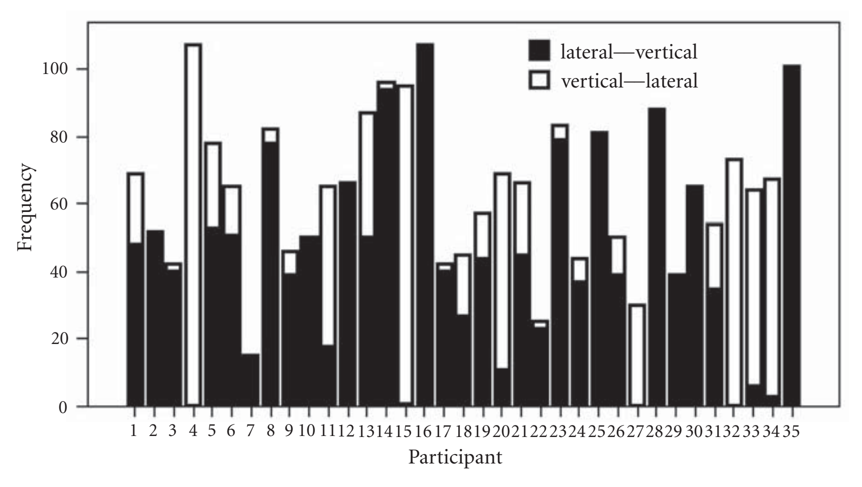


Fig. 4.7. Absolute number of utterances containing a syntactic combination in one of the two possible orders of dimension (lateral—vertical/sagittal vs. vertical/sagittal—lateral), per subject.

of dimensions over the other, as revealed by a Binomial test split by subject ($p < .001$ for 28 subjects, $p < .01$ for 4 subjects, and $p < .05$ for 1 subject). The distribution is shown in Figure 4.7.

These results show that—contrary to expectation—proximity to an axis is not the only factor determining order of spatial dimensions in a localizing utterance. At least for syntactic combinations, we observe a tendency to stick with one order of spatial dimensions across localizations (with a higher number of participants choosing the lateral-first order).

4.5 Discussion

The data presented here show that speakers tend to repeat spatial reference frame as well as lexical and syntactic choices across successive spatial utterances. Because of the inter-individual differences found, the consistency effects cannot be attributed to default options. Neither can they be put down to cognitive styles. First, it seems unlikely that we develop cognitive styles for using a particular word category to encode all directional relations. Second, importantly, the central finding, of a tendency to use the same kind of reference frame across trials, can be interpreted in terms of a *consistency effect* as it depends on participants' initial choice. This in turn can be influenced by manipulating the location of the object to be localized relative to the salient axes of the two reference frames in question. Similarly, the probability of using a prepositional vs. an adverbial pattern throughout a localization sequence is a function of initial lexical choice. This in turn is contingent on the typicality of a direction for a direction category, which again is determined by proximity to axes and edges (see Vorwerg and Rickheit, 1999, for 3D spatial categorization results; cf. Regier and Carlson, 2001, for an empirically founded attentional vector-sum model predicting acceptability judgements for *above, below, left, right* in 2D space, dependent on centre-of-mass and proximal orientations). Therefore, the data suggest that the combination of inter-individual variability and intra-individual stability can be accounted for by a consistency principle.

In methodological terms, consistency effects have to be taken into account as a possible factor affecting conceptual and linguistic choices in empirical studies with successive localizations. This may also concern practice trials. Different ways of conceptualizing spatial relations may be caused by randomization procedures providing different participants with different initial items. It is an open question to what extent a consistency principle may interfere with other factors aimed at in an experiment, possibly even with short-term interactive effects in dialogue.

Research by Watson, Pickering, and Branigan (2004, this volume) has shown that speakers are more likely to use an intrinsic frame of reference after their interlocutor used an intrinsic frame. This dialogue effect and the intra-speaker consistency effects described in this chapter may be based on the same kind of

representations activated by previous utterances. If the processes employed in language production and in language understanding draw basically upon the same representations (parity between comprehension and production), as suggested by Pickering and Garrod (2004), comparable post-conscious, automatic priming may occur from perceived and from produced previous utterances. Individual consistency in spatial language (with respect to frame of reference, word category used to denote spatial relations, and order of dimensions in combined spatial expressions) might therefore derive from the same cognitive mechanisms as the interactive-alignment processes in dialogue proposed by Pickering and Garrod.

On the other hand, it is also conceivable that speakers employ a deliberate consistency principle or strategy. This might be limited to the same type of spatial description (specifically, generalized direction terms, such as *left* or *behind*, as compared to cardinal directions; see Taylor and Tversky, 1996, for data on switching between survey and route descriptions). Speakers might intend to use the same lexical item (morpheme) in the same way—at least for the same kind of spatial configuration. Similar arguments have been put forward for perceptual judgements such as HEAVY and LIGHT, or LARGE, MEDIUM, and SMALL (Haubensak, 1992).

The present data do not precisely determine the cognitive processes leading to consistency. Future research is needed to assess to what extent speakers may be aware of their tendency to be consistent, or whether the results can be accounted for by automatic activation. Furthermore, if automatic priming is involved, its exact nature remains to be established (see Carlson-Radvansky and Jiang, 1998; Pickering and Garrod, 2004). However, it may be speculated—and the available evidence is consistent with this account—that automatic mechanisms of self-alignment between earlier and later representations during a discourse are at the heart of the observed consistency effects. This basic priming mechanism may be supplemented or compensated by deliberate decisions when a speaker becomes aware of reference frame choices by encountering difficulties (cf. Schober, this volume), verbal localizations made from different perspectives, or salient changes in the stimuli. However, in the absence of those factors, a pre-activation of underlying mental representations by initial selections may lead to consistency throughout a localization sequence regarding frame of reference, word category, and syntactic order. This is both the most parsimonious assumption and an account that easily relates results concerning reference frame choice and those concerning linguistic means.

5

An Interactionally Situated Analysis of What Prompts Shift in the Motion Verbs *Come* and *Go* in a Map Task

ANNA FILIPI and ROGER WALES

5.1 Introduction

While research that has looked at the question of how speakers go about instructing each other along a route is well established within an experimental paradigm, a growing body of work is now emerging that examines the co-construction of route-giving talk within a less controlled and more interactional framework (for example, Filipi and Wales, 2004; Muller and Prévot, this volume) and in naturally occurring interaction (for example, Kataoka, 2004; Psathas, 1986, 1991). Such a diversity of approaches is important if we are to elucidate what speakers actually do when they talk about space. As well, it is important to understand how cognition emerges in interaction (see Molder and Potter, 2005) and thereby avoid the couching of findings in ways that present a dichotomy between cognition and social understanding.

Complementary findings have emerged from these disparate approaches to route-giving or wayfinding talk. One that is pertinent to the current study is that speakers align or shift perspective or frame of reference (for example, Kataoka, 2004; Schober, 1993, this volume; Steels and Loetzsch, this volume; Taylor and Tversky, 1992a, 1992b, 1996; Tversky, 1996) and, from an interactional perspective, that they do so to facilitate the interaction by making their representations as coherent as possible (for example, Filipi and Wales, 2004; Kataoka 2004; Schober 1995, 1998) and to lessen the cognitive load (Steels and Loetzsch, this volume; Schober, this volume). Findings that have emerged from more interactional approaches have examined repair or communication breakdowns and the strategies speakers employ or need to employ to overcome or avoid such problems (for example, Anderson, 1995; Filipi and Wales, 2004; Shi and Tenbrink, this volume), conversational style including strategies to facilitate communication and coherence (for example, Anderson, 1995), studies on the structure of information (for

example, Filipi and Wales, 2003, 2004; Muller and Prévot, this volume; Psathas, 1986, 1991) and studies on the alignment of perspective shift to speakers' social stances (for example, Kataoka, 2004).

The present study builds on this work using the methods of Conversation Analysis—a micro-analytic, data-driven approach that provides a fine-grained analysis of (usually) naturally occurring talk. It provides a praxeological perspective on how language, cognition, and interaction may interrelate. Such a perspective places cognition in the orderly production of action in interaction as speakers work to co-produce talk using resources such as repair to deal with any problems that may arise (Schegloff, 1991, 1992). The analytic focus is on the moment-by-moment unfolding of talk with the next turn as the locus where speakers show how intersubjective understandings are achieved. This approach is not new to studies of space (see, for example, Psathas, 1986, 1991; Schegloff, 1972).

The study takes as its starting point findings reported in Filipi and Wales (2004) where we were interested in examining perspective taking and perspective shifting and the distribution of perspective categories within the sequence structure of the interactions. That work built on Taylor and Tversky's (1992a, 1992b, 1996) research on Route, Survey and Gaze perspective strategies. In sum, we found that the perspectives were sequentially and differentially distributed according to the interactional work that was being conducted.

We uncovered two route perspectives—a Route Internal and a Route External perspective—both of which were dynamic and associated with the task of carrying forward the action of direction giving. The Route Internal perspective located the addressee within the environment, within the world represented by the map; the Route External perspective located the speaker on the page which was external to the world represented by the map. The Survey perspective was linked to talk aimed at clarifying, confirming, or repairing any misunderstandings about the map, and the existence and location of landmarks, all important to the successful outcome of the task. As the speakers worked on the map task, they were found to shift their perspective throughout the interaction. We concluded that perspective shifting was motivated by a need to work cooperatively and collaboratively to get the task done, a finding aligned to Schober's (1995, 1998, this volume) conclusion that speakers switch their frames of reference when it is evident that their addressee is having trouble understanding them. Seen from this angle, perspective shifting can be construed as part of the process of alignment of perspectives between speakers when the spatial perspective chosen may be implicit. This alignment is something that speakers need to do in order to understand the instructions of their co-participant (Steels and Loetzsch, this volume). It thus becomes part of the speakers' repertoire for making spatial perspective explicit. In this chapter we return to the question of perspective shifting by focusing more pointedly on how shift is managed through the deictic verbs *come* and *go*.

These verbs are part of a larger set of motion verbs that are pervasive in interaction. They have been an object of intense study in linguistics. The most

influential theory of these verbs is Fillmore's (1971) theory based on the identification of 'appropriateness conditions' for deictically anchored linguistic expressions. Essentially, these verbs have been defined as reinforcing a locus for a point of view which encompasses a concept of motion and the passage of time from a starting point to an endpoint (Miller and Johnson-Laird, 1976). Accordingly, *come* indicates movement towards and *go* indicates movement away from a point of view, which itself is aligned to hearer, speaker, or the endpoint. The main driving force of these linguistically framed studies has been the search for default conditions for the deployment of these verbs.

Recently, a number of researchers have pointed to the gaps in the traditional distance parameter and a view of the indexical framework as fixed (for example, Cheshire, 1996; Cornish, 2001; Glover, 2000; Glover and Grundy, 1996; Hanks, 1992; Kataoka, 2004; Östman, 1995). Although the interests of these researchers differ, they share a consensus that it is necessary to ground analyses of deixis within a social context by examining speaker attitudes to both referent and addressee as speakers interact. These are features that have been marginalized in traditional geographical and temporal accounts of deixis (Cornish, 2001).

Hanks' (1992), Glover's (2000), Glover and Grundy's (1996), and Kataoka's (2004) analyses are particularly pertinent to our study. Their findings are based on a view that the origo shifts as speakers move through space and shift their position on and orientation to information and topics (Hanks, 1992). Glover (2000) and Glover and Grundy (1996) provide an analysis of how the choices and shifts in the indexical origo are systematically linked to the shifts in a speaker's perception of the problem at a given time. The shifts are shown to be sequentially significant and to serve a pragmatic politeness function. Within research on wayfinding, Kataoka (2004), in his study of the Japanese equivalents of *come* and *go*, argues that deictic verbs of motion provide both indexical cohesion (in a Hallidayan sense) and interactional cohesion which reflect speakers' emergent social stances to each other in the ongoing exchange.

By drawing on deixis from this interactionally situated framework, our ultimate goal is to examine how the speakers in our corpus align with each other and with the map as referent and how the verbs of motion (principally *come* and *go*) are indicative of these alignments. Specifically we are interested in what triggers a shift to and from the motion verbs *come* and *go*, how these are related to perspective taking and in which sequential environments the shifts occur. We are also interested in understanding what work is being done by the shifts in these verbs, whether they occur in children's interactions, and, if so, whether there are similarities with adults.

5.2 **Research Design**

The adult data on which this research is based were derived from the map task section of the Australian National Database of Spoken Language (ANDOSL) (Millar

et al., 1994). This corpus closely follows the Human Communication Research Centre (HCRC) Map Task (Anderson *et al.*, 1991) in design. Eight interactions of four adults comprise the adult corpus for our study. All were native speakers of (Australian) English.

The children's data were derived from the interactions of 16 children. Their ages ranged from 7;6 to 12;10. We grouped the children into two groups. In the younger group were eight children ranging in ages from 7;6 to 7;11. The second group was composed of eight children aged between 10;4 to 12;10.

The task required speakers to work in pairs using a map that the other could not see by assuming the roles of the instruction-giver (IG) or the instruction-follower (IF). The IG's map had a path marked on it. The IF had a similar map without a path. There were also differences between the maps with respect to the position, existence, and names of landmarks. The IG's role was to instruct the IF to draw the path onto her or his own map. Each pair performed in both speaker and listener roles using a different map each time.

There were some differences in the design of the maps and in procedures for the younger children in the corpus. Landmarks were included with which children were more familiar such as park and school; the finish (denoted by an 'X' and a picture) was marked on the map and a scenario for finding a trail left by a beach ball on the first map and a kite on the second was also created and read to the children. Prior to commencement of the task, we sat with the younger children individually to check that these landmarks were familiar to them by asking them to read the labels under the drawings. In addition to the general instructions to the children for doing the task prior to commencement, we also told them that on completion of the task they could exchange and talk about the completed maps. Recordings of the adults took place in a university laboratory. The children were both audio- and videotaped in one of the children's family home. There was no video data for the adults.

Full verbal transcripts were made of all adult route-giving interactions using the transcription conventions of Conversation Analysis (for example, Ochs, Schegloff, and Thompson, 1996: 461–5). In the case of the children's transcripts, non-verbal features were transcribed as well. Additional notations to capture these non-verbal features include:

{	which denotes the onset of a gestural feature,
→	to denote data of special interest,
EYF	to denote eyebrow flash,
- - -	to denote gaze engagement and
,,,	to denote gaze disengagement.

In keeping with the conventions of Language Acquisition studies, the age of the child in years and months appears in brackets next to the header at the start of each fragment.

5.3 Analysis

An analysis of the distribution of *come* and *go* yielded the following results.

- The default motion verb for both the children and adults was *go*.
- In two adult dyads' interactions there were frequent examples of shift.
- One adult dyad did not use *come*. The speakers adopted *go* and *head*.
- The fourth adult dyad only shifted to *come* once, at the end of the task.
- There were no examples of shifts to *come* in the younger children's interactions.
- There were six instances of a shift to *come* in two of older children dyads' interactions; as a comparison, the total number of instances of *go* produced by these two dyads was 46 (23 per dyad).

We need to go beyond the count analysis, however, if we are to understand what work is being achieved by each of these verbs as the IG and IF interact. We begin by analysing verb shift as being aligned to perspective shift in the adult interactions.

5.3.1 Adult data: Shifts in *come* and *go* aligned to shifts in spatial perspective

Analysis of the adult data reveals that a shift in verb choice is directly aligned to the shift in spatial perspective. This is illustrated in the following fragment.

Fragment 1
Dyad 3 (Adult)

1	IG:	→	. . . and then underneath the words consumer trade fair. just
2			come underneath that okay?
3	IF:		[yeah,]
4	IG:	→	[AND] THEN go due slightly north of west (0.5) so you head
5			back (0.7) uh::: towards the left-hand side of the page but just
6			lift your track up a little bit (0.4) so that you n- y- you just
7			about clip (0.3) the 'c' in consumer (0.4) awright¿
8	IF:	→	the what?
9	IG:		the 'c' in [consumer]
10	IF:		[oh r]ight=
11	IG:	→	= ◦trade fair awright¿◦(0.4) and then you come across, (0.4) what's
12			annoying is that you don't have this river on your map to give me-
13			give you a bearing but. . . .

On one level, in a more traditionally deictic sense, we can see that the verbs are associated with the position in which the speaker places himself and his co-participant conceptually on the map. The shifts are working on another level as well. The IG shifts his perspective to make the instructions more explicit for the IF. This is evident in the shifts in lines 1 and 4, from *come underneath* (line 1) to go *due slightly North*. He has thus moved from more specific instruction and details

in relation to the page or landmark using the Route External (*then underneath the words consumer trade fair just come underneath that*) to more general information about direction contained in the Route Internal (*go due slightly north*). We note further that the switch to *come* is aligned to a familiar landmark—*the consumer trade fair*—which has come into the established, shared information about the map discovered through the talk so far. The IG then shifts back to a Route External perspective by making reference to the page. In the shift to *come* in line 11, he places himself ahead of the IF on the map and therefore at the endpoint—*then you come across*. At this point the instruction is suspended as the IG looks for a landmark or feature that is common to both maps as an anchor. Thus, as well as the conventional view that these verbs encode a physical space, we can see how the shifts are determined by interactional ends. In keeping with Kataoka's (2004) claim, these verbs encode both a spatio-physical and socio-cognitive configuration.

While the verbs *come* and *go* here are aligned to shifts in spatial perspective and the achievement of different kinds of information about the direction in which to move along a path (and to that extent are working as topic boundary markers: Kataoka, 2004), the discourse marker *so* can also be seen to be playing a pivotal role to linguistically mark a shift in the information, although it does not necessarily mark a spatial shift in the way that the verbs *come* and *go* do. As Muller and Prévot (this volume) maintain, discourse markers and response tokens play an important role in understanding how information is grounded between speakers. They have also been shown to signal perspective shift or maintenance (Filipi and Wales, 2003). Here, the discourse marker *so* can be seen to be working to ground the instruction by reformulating or expanding the information in the preceding utterance, the purpose of which is to provide more explicit instructions for the IF.

There are a number of examples in the data where shifts in *come* and *go* are associated with the shifts from or to the Route External or Route Internal perspective. The static verbs *be* and *have* on the other hand are associated with the Survey perspective. The next fragment illustrates.

Fragment 2
Dyad 3 (Adult)
1 IG: right well just [head]
2 IF: [ahh]
3 IG: → head south of your dingo open-cut mine.
4 IF: → am i to [the-]
5 IG: → [come] straight down.
6 IF: am i to the left-hand side or the right-hand side of the d- of
7 your gala open-cut min[e],
8 IG: [l]ooking at it you're on the left.
9 IF: okay so i just come straight down from [the cross to the-]
10 IG: [come straight down] the left.

The IG reformulates her instruction as she shifts from the dynamic verbs *head south* (used as an equivalent to *go*) to *come straight down*. The effect is to conceptually situate the IF at the starting point and herself at some point further along the map on the page. The shift appears to be triggered by the interrupted action of the IF in line 4. We note his shift to the static verb to be (*am I*) in line 4 which marks a switch to the Survey perspective in line 6 as he attempts to seek clarification about his orientation with respect to the landmark. In the IF's next utterance in line 9, there is a shift in perspective back to Route External as he aligns with the IG by using *come* in his repeat of the IG's original instruction. The alignment of the speakers is also evident in their use of their addressee's perspective in the IG's use of *your dingo open-cut mine* in line 3 and the IF's use of *your galah open-cut mine* in lines 6–7 (in opposition to their own open-cut mines). This accords with Schober's (1995) finding that speakers use their addressee's perspective as a way of accommodating and facilitating comprehension. This fragment provides a nice example of how the shift in talk about the map to route instructions is aligned to and achieved through shifts in the choice of verbs, and how this in turn acts as a facilitative strategy.

5.3.2 Triggers

There are several triggers that prompt the switch in verbs. One is a reformulation which re-does the initial instruction. Another is an expansion of the instruction. Both of these are present in the above two fragments. These 'redoings' and expansions are common throughout the data as the IG works on providing clear sets of instructions. Self or other initiated repairs also act as triggers. An example is provided in lines 7 to 9 in the next fragment where the IF initiates repair and draws attention to what he perceives to be the IG's misleading instruction.

Fragment 3
Dyad 3 (Adult)
1 IG: → [s]o bring m∼ so you've now∼ after the millionaire's castle
2 you've come down about three inches,
3 IF: → mm hm¿
4 (0.4)
5 IG: → and you do a big sweep up to your right?
6 (1.1)
7 IF: → ◦okay◦. sweep up? joanie if i'm going to my right i can't go up i
8 mean right is (0.4) east (0.3) or (0.5) east is towards the le- the
9 right hand side of the page.=
10 IG: =yeah,
11 (0.4)
12 IF: and north would be up towards the top of the [page]
13 IG: [that's right so
14 you're] swee [ping east.]
15 IF: → [so am i sweep]ing
16 right around [am i going east?]

17 IG: [you're sweeping] east.=
18 IF: → =yeah okay well don't- yeah okay all right i'm going east (0.3)
19 i'm going all right >under my ghost town< (0.3) then i've coming to a ROLLING
20 stone creek¿

Initially, the IF is taking issue with the IG's use of *up*, in the cross-over between *up* and *east* in the IG's instruction. As they deal with this issue in talk about the spatial terms, the direction-giving is momentarily suspended at lines 7 to 20. In the IF's turns (lines 15 and 18 to 20), we note shifts in spatial perspective and verb as he moves to Route Internal (*I'm going east*) and then to a Route External *then I've coming to a rolling stone creek* to designate the endpoint. The shift in line 18 from *I'm going under my ghost town* to *I've coming to a rolling stone creek* is a description of the line that the IF is drawing and an indication of arrival at a next landmark. Its prosodic contour and its placement in the utterance suggest that it is being offered to the IG for confirmation. Interestingly, the switch here occurs in the IF's utterance. Given the roles of the IF and the IG and their differential access to information about the path, this is an infrequent action by the Information Followers in our corpus and only occurs in this dyad's interactions.[1]

Expansions, reformulations and repair can thus trigger a shift in perspective. They are resources used by the speakers to create a clearer set of instructions. However, while reformulations and expansions are used much more frequently by the IG, who is privy to more information about the path than the IF, repair is available to both the IG and the IF. 'Other'- initiated repair in particular is a resource used by the IF when an instruction is unclear or when checking that an instruction has been understood.

5.3.3 Shifting to *come* as a signal of grounding of information as jointly shared

So far we have examined sequences where the shift in verb choice is directly aligned to the shift in spatial perspective and appears to be working to differentiate between established information about the maps (known landmarks) and unknown information (landmarks to be established as common to both maps). In the next two fragments, we look more closely at how the shift to *come* signals that this information becomes grounded between speakers.

[1] Indeed, all the instances of verb shifts in this IF's utterances similarly carry this function of offering the IG a report of where he has come, of what he has done, or of what he is about to do. The verb tenses are therefore either in the past, future, or present continuous. In contrast, with the exception of the shifts which are working to summarize the path taken so far expressed in the past tense, the IG's shifts occur as directives in the action of direction-giving.

Fragment 4 occurs in the concluding phase of the map task.

Fragment 4
Dyad 4 (Adult)

1 IG: [yes now] (0.5) is there a thing called a statistics
2 [centre ?]
3 IF: [yes] ye[s,]
4 IG: → [well] head between the cotton fields [and the] statistics
5 centre,=
6 IF: [yes]
7 =yes
8 (0.4)
9 IG: → and uh and then come down towards the old deserted lighthouse=
10 IF: =yes that's- (0.5) [well]
11 IG: [and when you've] reached that [you've uh]
12 IF: [yes i've got to
13 the old] lighthouse, well that's where the treasure is that's where we
14 [stop].

The fragment opens with perspective-free talk involving a question and answer sequence to establish whether the statistics centre is a common landmark on both maps. There is a resumption of the instruction giving in line 4—*well head between the cotton fields and the statistics centre*. In line 9, the IG places himself at the endpoint and directs the IF to come towards the final landmark on the map—the deserted lighthouse. This is an interesting shift coming as it does right at the end because it signals that the IF now shares the IG's information about the path and the landmarks which have become part of their shared knowledge. By extension, the roles of IG and IF are no longer relevant. It marks finality, a conclusion to the task.

This association of *come* with completion of the description of the route taken or path drawn, and therefore with the landmarks becoming established as shared knowledge, occurs within the body of the task as well, as displayed in the following fragment.

Fragment 5
Dyad 3 (Adult)

1 IG: =◦oh there's nothing there okay.◦ (0.2) awright. (0.8) now when you
2 get um not quite you're about an inch above the statistics centre, (0.2)
3 okay¿ (0.1) p- an- and from about the middle of the statistics centre so
4 take a bearing from the middle of the stis- the statistics centre¿
5 IF: yeah,
6 (0.5)
7 IG: → and come up about an inch (0.4) and you're in a- you co- you've come
8 over- you've come away from the coast. (0.1) what i want you to do then
9 is diagonally go southwest (0.6) go southwest, (0.6) due southwest, so in

10 a forty-five degree line just go straight down towards the (0.4) western
11 margin, (0.4) [okay¿]
12 IF: [right,]

After some lengthy talk about finding a common landmark which is on the IG's map but not on the IF's, the perspective shifts to Route External in line 7 (*come up about an inch*). The shift works to locate the IF on the place that is marked by the landmark on the IG's map but an empty space on the IF's. The IG's task is thus to direct the IF to a position common to both maps. It is essential to establish this position in order to proceed to the next place on the map. The choice of *come* serves to direct the IF to the IG's position. It marks the status of the information just described as now shared by both the IG and the IF. It therefore enters their common, established information about the map. The shift to *go* in lines 9–10 (*diagonally go ... just go straight down*) marks a shift to new information yet to be divulged or negotiated which is 'owned' by the IG in her capacity as Instruction Giver. It marks movement towards a next location and therefore a resumption of the direction-giving as opposed to the action of suspension of direction-giving in order to find a bearing.

This same framing of information as shared and information as new is achieved by pronoun shift rather than verb shift in the interactions of the dyad that only uses *go*.

Fragment 6
Dyad 1 (Adult)
1 IG: → yep (0.5) okay now we've just crossed the stream under the spruce
2 tree[s]
3 IF: [yeah]
4 (0.7)
5 IG: follow the stream the bend of the stream but about quarter of an
6 inch out,=
7 IF: =yeah,
8 (0.3)
9 IG: → um (0.5) till you get near the top of the consumer trade fair.
10 IF: okay [i've done that.]
11 IG: → [and then] (1.0) go (0.4) east (0.4) and around and
12 underneath the consumer trade fair.

Here we can see the effect of the shifts in pronouns from the inclusive *we've* to denote the joint position of IG and IF on the map (*we've just crossed the stream* in line 1) to the 'other'-centric *you* (*follow the stream*) in line 5, and (*till you get near the top*) in line 9. The shifts in pronoun and tense work in similar ways to the shift from *come* to *go*. The juxtaposition of these two examples and their sequential analysis makes it possible to see that a different set of linguistic resources is being used to do the same work.

5.3.4 Children's data

As already noted, our analysis yielded six occurrences of *come* in the children's data. All occurred in the interactions of two of the older children's dyads. Four occurrences were produced by dyad 5. All occurred in environments where a landmark was absent from the IF's map and the speakers were compelled to try to establish a common position before direction giving could resume (as in the adult fragments 4 and 5 just analysed). The next fragment, where a shift to *come* is found at the end of the direction-giving, illustrates.

Fragment 7
Dyad 5 (Children 11;7 and 11;0)
1 IG: then go about 4 centimetres to the ↑left, (1.1) then {go:: (1.1)
{- - - IF , , ,
2 .hh huh ((sighs)) (0.3) sou {::th ↑{we::::{st ∘or: (0.4) left dow::n
{-→IF{EYF
{IF ⟶ IG
{((IG Moves RH down then
up to the right then down again.))
3 → ( )∘{(0.9) for about 6 ↑ centimetres, >you sort ov< come to the:
{, , ,
4 (.)∘↑ si::de. ∘(1.3) go down a couple ov centimetres from there (0. 5)
5 and then go to the ↑ right. (.) ∘and put a cross. (0.3) that's the end.∘

Here the shift occurs in an expansion of the instruction in line 3 after a shift in spatial perspective (line 2). In the absence of specific landmarks, the IG is attempting to give further helpful information to the IF by once again drawing on a feature that both participants have in common—the side of the page. *Come* is the choice of verb in this context where a feature that is common to both maps is being referred to and used as an anchor. The association of *come* with shared or established landmarks and features is once again evident.

There are two instances of a verb switch in the second child dyad's interactions. One appears in a reformulation introduced by the marker *so* and mirrors the switch noted in the adult sample in fragment 1. The other instance (analysed below) achieves very specific work.

Fragment 8
Dyad 9 (Children 12;3 and 10;4)
1 IG: from there (0.3) do you have anything to the diagonal down left?
2 (0.2)
3 IF: {down left?
{((Places hand on page.))
4 IG: diagonal down {left,
{—→ IG
5 IF: the millionaire's {castle.
{((Hand goes up to her face.))

6 IG: → {YES:::!( ) i have it (too)! {now what you do is come around to the
{((Gestures with closed fist in victory gesture.))
{((Both look down at the page.))
7 the top of the millionaire's castle,
8 IF: yeah,
9 IG: and then start down into the middle of the millionaire's castle, . . .

At 21 minutes, this dyad's interaction is decidedly long when compared to an average of four minutes for the other children. The first occurrence of *come* appears 12 minutes into the interaction. Up until this point, the speakers have had enormous trouble in finding shared landmarks. They have also had difficulty in working out the best way of dealing with the problem of navigating around a map when landmarks are not shared. In fragment 8 *come* is produced at precisely a point where the speakers have come out of a protracted process of finding a way around absent and different landmarks on each other's maps by finally discovering a landmark in common. The IG's reaction to this discovery is displayed in line 6 both verbally and in gesture (*YES! I have it too!* accompanied by a gesture of victory). The shift from the Survey perspective to the Route External perspective in line 6 is marked by *come.* It seems to be working to mark arrival at a common point on the map from which the directions can then proceed.

5.4 **Discussion**

Not all the speakers in our corpus shifted to *come*. As we have seen, there were very few examples of verb shift in the older children's interactions and no instances of shift were recorded in the interactions of the younger children. This is despite the fact that they were able to shift perspective from Route to Survey. How might we account for the absence of *come* when developmentally we know that children are able to shift perspective from a relatively young age provided certain conditions are met (Ziegler, Mitchell, and Currie, 2005)? One reason we propose for the absence of verb shift is that the task of giving instructions is itself cognitively demanding. Considering the work that is being accomplished by the shift in verbs, the processes involved and reasons for the shift, any further shifting would increase the cognitive load. This concurs with Anderson's (1995) findings. Using a map task with children of a similar age group to the children in our study, Anderson concludes that the children do less grounding work to establish mutual knowledge than the adults. She proposes that a reason for this may be that the process is a demanding one and that the children's cognitive resources are garnered to formulating instructions.

In the interactions of the speakers where verb shift was present, it was aligned to the shifts in spatial perspective and to the speaker's stance to the co-participant with respect to information owned and shared. *Come* was aligned with the Route External perspective and *go* with the Route Internal perspective. In adopting the

Route External perspective, speakers made reference to specific landmarks, to the drawing of the path and to features on the page such as distances of the line between landmarks. In adopting the Route Internal perspective, speakers made reference to more general movement through space, frequently using the cardinal points. The shifts were triggered by self (in reformulations, expansions, or self-repair) or by the other speaker (manifested structurally through repair initiations).

In addition to the alignment of verb shift with spatial perspective shift, speakers also shifted to *come* when describing the route taken so far or to mark task completion. By shifting to *come*, the IG seemed to be marking points in the task as shared arrival points. Speakers in our corpus who did not mark their spatial talk with shifts between *come* and *go* used other linguistic devices (pronoun shift for example) to achieve the same end. This sense of shared arrival also underlies the use of *come* in the interactions of the older children, for whom the switch to *come* was rare and very much associated with resolution of a preceding problem when faced with the absence of shared landmarks. Verb-shifting can thus be viewed as being implicated in the process of grounding information between speakers very much like the response tokens and discourse markers described by Muller and Prévot (this volume). These do very specific and differential work in grounding, accepting, and anchoring information.

In trying to account for these shifts, we can draw on the traditional distance parameter and the concepts of proximity and distalness (Fillmore, 1971) by linking these concepts to the speakers' stance to her/his addressee, to the task and to her/his role as IG or IF. The first link we can make is the association of the Route Internal perspective with the distal *go* because it is associated with movement through a world that is conceptualized by the speaker and is therefore distant from the co-participant because it is not yet shared. It becomes shared through interaction. Concomitantly, the Route External perspective is associated with the proximal *come* because it is aligned directly to the physical map on the page—with the map that both speakers can see and share despite some differences in landmarks. It is therefore associated with proximity in terms of what is familiar, established and shared. This notion of proximity also underlies the other uses of *come* which our analysis has uncovered—the conclusion of the task and the shift in ownership of the information as jointly owned, the resolution of a problem (potential or real) by finding a shared landmark and the summary of the path taken so far as a shared achievement before setting off on new ground, all of which are achieved through interaction.

In each instance, the origo is not fixed, but shifts as speakers move through space, shift their perspective and their position on and orientation to information which they must share in order to complete the map task efficiently and cooperatively. Also implicated here is speaker perception of his/her role and the stance taken to the other speaker within the roles assigned to them. This may explain why some speakers adopt *go* throughout (perhaps taking a stance to the

task throughout as the 'owners' of information, which is aligned to Kataoka's (2004) conclusion with respect to social stance), while others shift between the verbs as their claim on information shifts with the ongoing step-by-step completion of the map.

5.5 Conclusion

This analysis has attempted to show that the traditional deictic concepts that underlie *come* and *go* are enriched when the analytic interest shifts to a focus on interaction and to perspective taking. The perspectives taken by the speakers are fluid and serve the function of minimizing the effort required to get the task done. The speakers' roles are dynamic as well. The IG may have most of the information necessary to complete the task, but the IF also has a role to play in providing information to the IG. All this is achieved within the moment-by-moment shaping of the talk and has to be seen as the joint accomplishment of the speakers who share perspectives or shift perspectives, as marked by the changes in verb choice. As Schober (1998) maintains, it is the coordination of multiple perspectives in conversation that is important for the success of communication.

The shifts in verbs are thus to be seen within a framework that views deixis as socially situated, as occurring in an interactional space (Hanks, 1990; Hindmarsh and Heath, 2000). These are the products of both speakers who themselves shift from speaker to hearer and whose actions constitute a set of resources for getting the task done.

6

Perspective Alignment in Spatial Language

LUC STEELS and MARTIN LOETZSCH

It is well known that perspective alignment plays a major role in the planning and interpretation of spatial language. In order to understand the role of perspective alignment and the cognitive processes involved, we have made precise complete cognitive models of situated embodied agents that self-organise a communication system for dialoging about the position and movement of real world objects in their immediate surroundings. We show in a series of robotic experiments which cognitive mechanisms are necessary and sufficient to achieve successful spatial language and why and how perspective alignment can take place, either implicitly or based on explicit marking.

Spatial language consists of expressions that involve spatial positions and movements of objects in the world. Spatial language often involves perspective (Schober, 1993, and this volume). For example, the meaning of the phrase 'the ball left of the glass' depends on the spatial position of the viewer with respect to the objects involved. Moreover, this viewer can be the speaker (egocentric) or the hearer or somebody else involved in the conversation (allocentric). In any case, if speaker and hearer perceive a scene from different perspectives, they need to align the perspective from which the scene is being described in order to make sense of the description. Often perspective is implicit and dialogue partners must then indirectly align perspective. But natural languages also have various ways to make perspective explicit, as in 'the ball to my left' (see also Carlson and Hill, this volume).

The goal of our work is to explain these well-known facts. Concretely, we would like to understand why perspective is unavoidable in spatial language, how dialogue partners can still align perspective even if it is not marked, and why and how marking helps. We would also like to understand how the whole system can get off the ground—in other words how spatial language involving implicit or explicit perspective alignment can be learned or invented through negotiation in consecutive dialogues.

Our explanations will be based on making very precise and complete models of communicating embodied agents, situated in a particular real-world environment. The models are complete in the sense that they include mechanisms for

achieving physical behaviour in the real world, vision for the construction of situation models, cognitive mechanisms for developing and using spatial categories like left/right, forwards/backwards, close/far, and mechanisms for developing and using lexicons. Our models have been completely formalized and implemented on physical robots so that we can test their effectiveness and behaviour in repeatable experiments. In each experiment, we set the agents up to play situated language games in the form of dialogues about the objects in their world. The agents describe to each other the movement of a ball in their close proximity. Because spatial language is obviously a very useful and effective way to do so, we expect it to emerge as part of consecutive games.

We will make three arguments:

1. *As soon as agents are embodied, they necessarily have a specific view on the world and spatial language becomes impossible without considering perspective.* We will show this by an experiment in which first the agents see the world through the same camera (in other words two agents use the same robot body) and hence they have exactly the same visually derived situation model. And second, the agents are made to see the world through their own camera and so they each have a different situation model. The experiment clearly shows that in the second case, a communication system cannot get off the ground. They do not learn the meaning of spatial terms from each other and generally fail to understand each other.
2. *Perspective alignment is possible when the agents are endowed with two abilities: (i) to see where the other one is located, and (ii) to perform a geometric transformation known as the Egocentric Perspective Transform.* This transformation allows the agent to compute what the scene looks like from the viewpoint of the other, in other words to develop a situation model from the other partner's perspective. The Egocentric Perspective Transform is normally carried out in the parietal-temporal-occipital junction (Zacks et al., 1999) and used for a wide variety of non-linguistic tasks, such as prediction of the behaviour of others or navigation (Iachini and Logie, 2003). We have implemented these capabilities and performed an experiment in which agents test systematically from which perspective an utterance makes sense. They are thus able to implicitly align perspective, but only because they are both grounded and situated in the same real-world setting. The experiment demonstrates that agents are in this case able to bootstrap spatial language and achieve successful communication. Note that this is still without explicitly marking perspective.
3. *Perspective alignment takes less cognitive effort if perspective is marked.* We investigate this through another experiment that compares the implicit way of perspective alignment (as in 2.) with one where perspective becomes marked because the lexical processes now express to what perspective the speaker/hearer is aligned (egocentric or allocentric). We observe a significant decrease of cognitive effort. This experiment shows additionally that our models are adequate

for demonstrating how perspective markers can be invented and learned. This is not a simple problem and children can only do it fairly late in language development.

The remainder of the chapter is in two main parts. The first part (6.1) gives more details on the experimental setup and on the various cognitive mechanisms that make up the agent architecture. The second part (6.2) reports results of our experiments. A final section (6.3) derives some conclusions.

6.1 Experimental Setup

A lot of work has recently been done on studying human dialogue (Clark, 1996; Pickering and Garrod, 2004; and several contributions to this volume). The methodological approach discussed here is entirely complementary. We take the findings of these investigations as given but try to see what it takes to build synthetic models of dialogue, which obviously requires a 'mechanistic' theory of all the processes involved in dialogue and a concrete setup where we can test these processes. Moreover, we are interested in understanding how spatial language with perspective marking can arise in a population, motivated by attempts to understand the origins and evolution of communication systems (Steels, 2003).

Our experiment uses physical robotic 'agents' which roam around freely in an unconstrained indoor environment (see Figure 6.1). The agents have subsystems

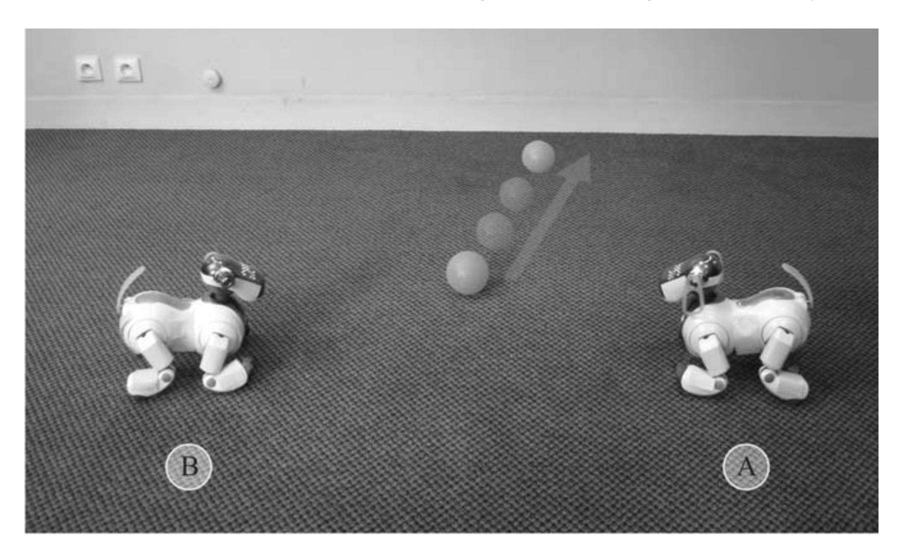


Fig. 6.1. Agents embodied in physical robots. The speaker (in robot A) and the hearer (in robot B) together observe ball movement events and then play a language game to describe the scene to each other.

for autonomous locomotion and vision-based obstacle avoidance. They maintain a real-time analogue model of their immediate surroundings based on visual input (see Figure 6.2). Using this analogue model, the robots track other robots as well as orange balls using standard image-processing algorithms. Furthermore, the robots have been endowed with a subsystem to segment the flow of data into distinct events and they then build a situation model. There is a short-term memory which contains the situation model of the most recent event and a number of past events.

The robot agents engage in language games—routinized communicative interactions. Two robots walk around randomly. As soon as one sees the ball, it comes to a stop and searches for the other robot, which also looks for the ball and will stop when it sees it. Then the human experimenter pushes the ball with a stick so that it rolls a short distance, for example from the left of one robot to its right. This movement is tracked and analysed by both robots and each uses the resulting perception as the basis for playing the language game, in which one of the two (henceforth the 'speaker') describes the ball-moving event to the other (the 'hearer'). To do this, the speaker must first conceptualize the event in terms of a set of categories that distinguishes the latest event from previous ones, for example that the ball rolled away from the speaker and to the right, as opposed to towards the speaker, or, away from the speaker but to the left. The speaker then expresses this conceptualization using whatever linguistic resources in his inventory cover the conceptualization best and have been most successful in the past. The game is a success if, according to the hearer, the description given by the speaker not only fits with the scene as perceived by him but is also distinctive with respect to previous scenes.

Agents take turns playing speaker and hearer so that they each gradually develop the competence to speak as well as interpret. No prior language or prior set of perceptually grounded categories is programmed into the agents. Indeed the purpose of the experiment is to see what kinds of categories and linguistic constructions will emerge, and, more specifically, whether they involve perspective marking or not.

The agents use two additional subsystems to achieve this, as described in more detail shortly. The first one performs categorization and category formation (Harnad, 1987). We use here discrimination trees (as explained below), although other categorization methods (e.g. Radial Basis Function networks or Nearest Neighbour Classification) would work equally well. The agents apply categorization to the sensory channels that directly reflect properties of the visual image computed using standard image-processing algorithms, such as start and end position of the ball, angle of the trajectory, distance travelled by the ball, etc. The second subsystem concerns the lexicon. We use a bi-directional associative memory which associates one pattern (here a set of categories) with another pattern (here a word). The associations are weighted with a score because the same pattern may be associated (in either direction) with more than one other pattern. Indeed, one word can have many meanings (polysemy) and several words

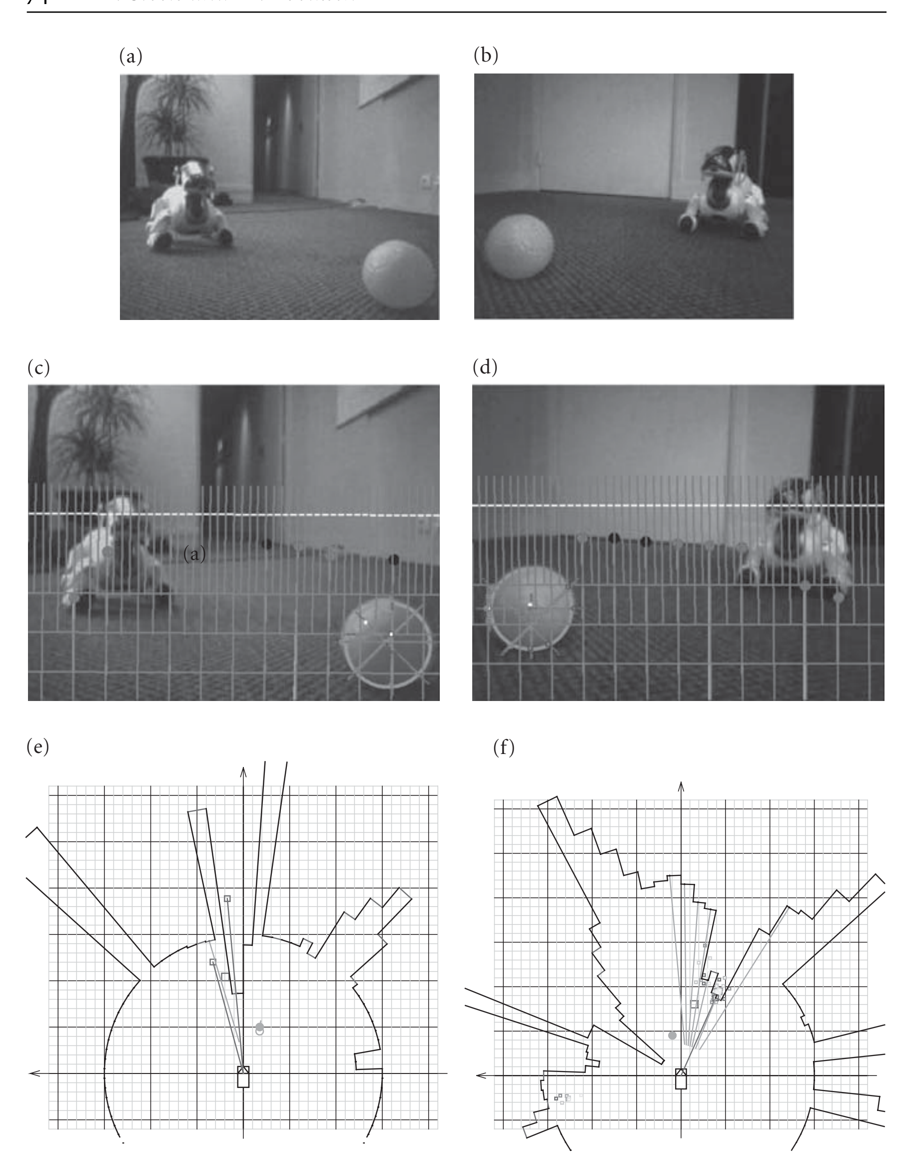


FIG. 6.2. Top row: the scene from Figure 6.1 seen through the cameras of robots A and B. Second row: from each image, the positions of the ball, other agents, and obstacles are extracted. The images are scanned along lines orthogonal to the horizon for characteristic gradients in the colour channels. Bottom row: the agents maintain a continuous analogue model of their immediate surroundings by integrating the (noisy) information extracted from the camera images. The graphs show snapshots of this model at the time when the images in (a) and (b) were taken.

can be in competition for the same meaning. In retrieving a target given a source, the association with the highest score is preferred. Neural implementations of bi-directional associative memories have been well studied and shown to be applicable in a wide range of domains (Kosko, 1988).

The behaviour of the two subsystems (for categorization and lexicon lookup) is structurally coupled in that success in the game raises the score both of the categories that were used and of the lexical conventions that were used to express those categories, so that agents progressively come to share not only their linguistic conventions but also their conceptual repertoires (as extensively shown in Steels and Belpaeme, 2005).

In addition to subsystems for visually perceiving and acting in a dynamically changing world, extracting and memorizing events, discriminating events from previous ones using discrimination trees, and lexicalizing these distinctions using a bi-directional associative memory, agents are endowed with a subsystem for egocentric perspective transformation so that they can reconstruct a scene from the viewpoint of another agent. This requires that they first detect where the other agent is located (according to their own perception of the world) and then perform a geometric transformation of their own world model. Inevitably, an agent's reconstruction of how another agent sees the world will never be completely accurate, and may even be grossly incorrect due to unavoidable misperceptions both of the other robot's position and of the real world itself. The sensory values obtained by the robots should not be interpreted as exact measures (which would be impossible on physical robots using real-world perception) but at best as reasonable estimates. This type of inaccuracy is precisely what a viable communication system must be able to cope with and robotic models are therefore the only way to seriously test and compare strategies and the mechanisms that implement them.

The following subsections provide some more technical detail and examples of each of these subsystems at work. Readers who are not interested can skip the remainder of this section and immediately look at the results of the experiments on perspective alignment and perspective marking.

6.1.1 Embodiment, behaviour, and perception

As an experimental platform we use the Sony ERS7 AIBO, which is a highly complex fully autonomous and fully programmable robot. In addition to the on-board computing power, we use an external computer to control the experiment and engage in some of the symbolic aspects of each robot's behaviour. Although there has been a lot of progress in robotics during the last few years, particularly due to the rise of the 'behaviour-based approach to robotics' (Steels and Brooks, 1994), doing perception and autonomous behaviour with real robots is still an extremely difficult task. We could not have done this experiment without relying on the existing robot soccer software developed by Röfer *et al.* (2004). The vision

system has to deal with noisy and low resolution (208 x 160 pixel) images from a robot's camera. Objects like the ball look very different in different places of the environment due to slight differences in illumination. Noisy perception introduces the challenge of maintaining a robust situation model. As the perception cannot always be trusted, the resulting position of the ball is only an estimated position gained with probabilistic filtering techniques. As shown in Figure 6.3, the two robots never perceive the scene in exactly the same way.

Behaviour-based control systems (Loetzsch *et al.*, 2006) were implemented for the physical coordination of the robots. Both robots randomly walk around while avoiding obstacles. Each robot that sees both the ball and the other robot sends an acoustic signal. Robots continue with random exploration until a configuration is reached such that they both see the ball and the other robot and know that the other robot is doing so as well (that is, they establish a joint attentional frame in the sense of Tomasello, 1995). When both robots are ready to observe the scene together, a human experimenter manually moves the ball. The beginning and end points of the trajectory are recorded and sent to the language system via a wireless network (see Figure 6.3).

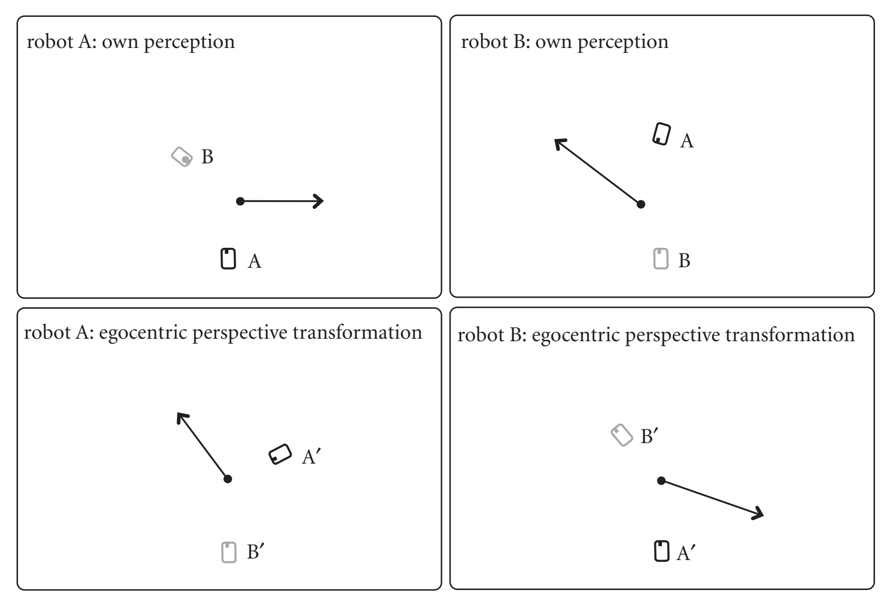


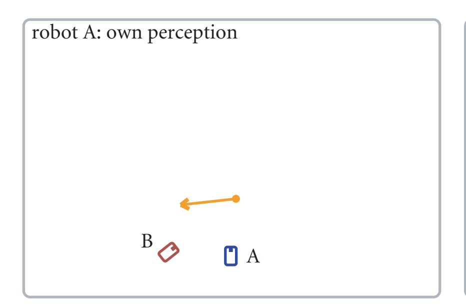
Fig. 6.3. The agents are endowed with the ability to segment the continuous stream of visual data (Figure 6.2) into discrete event descriptions that make up their situation model. Top row: the event from Figure 6.1 as perceived from robots A and B. Bottom row: the result of egocentric perspective transformation. Both robots are able to construct a description of the scene as it would look from the perceived position of the other robot.

As shown in the bottom row of Figure 6.3, each robot is able to compute an additional description of the scene from the perspective of the other robot (egocentric perspective transform) so that they are in fact able to compute the situation model from a perspective other than their own. Note that this situation model is not always accurate (due to the difficulty for each robot of perceiving the perception of the other). In Figure 6.3 robot A's situation model of B (bottom left in Figure 6.3) is slightly different from robot B's actual situation model (top right in Figure 6.3).

6.1.2 Conceptualization by the speaker

The goal of the conceptualization subsystem is to come up with the meaning to be expressed by the speaker. This meaning should be such that it discriminates the topic (the most recent event) from the other events in the context. Conceptualization decomposes into three subsystems. The first one extracts a battery of features from the perceived scene. The second categorizes the objects in the context based on these features, and the third subsystem finds out which categories are discriminative. Because the speaker can compute the scene from the perspective of the hearer, he will not only conceptualize from his own perspective but also from that of the hearer so that he can determine whether perspective needs to be marked or whether he is going to be more successful describing the scene from the perspective of the hearer because that is more salient and can be done with more established categories.

It is helpful to see the operation of the different subsystems for a concrete example. We take the 4116th interaction from a series in a population of five agents. Agents 3 and 4 were randomly drawn from the population; agent 3 was randomly assigned to be the speaker and 'used' robot body A. Agent 4 was the hearer (robot B). Both have perceived two events (for robot A shown in Figure 6.4).



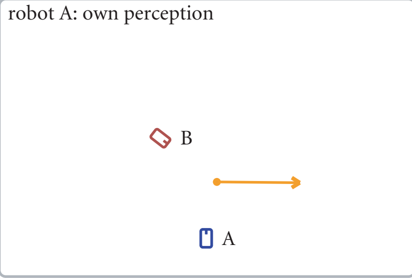


Fig. 6.4. Two events as subsequently perceived by robot A. The goal of conceptualization is to find a set of categories that discriminate the recent event (b) from the previous event (a).

channel	description	value	scaled
ball-x1	x of start point	477.28	0.41
ball-d1	distance to start point	489.14	0.42
ball-x2	x of endpoint	469.21	0.34
ball-y2	y of endpoint	−941.97	0.25
ball-d2	distance to endpoint	1052.37	0.47
ball-a2	angle to endpoint	−63.52	0.17
roll-angle	angle of movement	−90.55	0.18
roll-distance	length of trajectory	834.95	0.47
delta-x	change in x	−8.07	0.37
delta-y	change in y	−834.91	0.28
delta-a	change in angle	−50.87	0.22
delta-d	change in distance	563.21	0.47

Fig. 6.5. The feature values for the event in Figure 6.4a

Categorization operates over 12 feature channels which are calculated for each event based on straightforward signal processing and pattern recognition algorithms (see Figure 6.5). For example, channel ball-x1 is the x component of the start position of the ball, ball-y2 is the y position at the end of the movement, delta-a is the change in angle of the ball, and so on. For ease in further processing and in order to be able to compare features, each channel value is scaled within the interval [0 . . . 1]. 1 means that it is a very high channel value (with respect to the observed distribution for that particular channel) and 0 a very low value.

Categorization itself is performed with a discrimination tree approach described in more detail in Steels (1996). In order to help the hearer guess what the speaker meant, the most *salient* feature is chosen. Saliency is computed as the minimum distance of the feature values of the topic to the average feature values of other events in the context:

channel	ball-y2	delta-y	roll-angle	. . .	ball-x1	ball-d1
saliency	0.72	0.70	0.52	. . .	0.00	0.00

As shown in Figure 6.4, the features ball-y2 (end position left/right) and delta-y (change towards left/right) are much more salient than ball-x1 (start position far/close) and ball-d1 (distance to the ball at the beginning).

There is a discrimination tree for every feature channel. Each tree divides the range of possible values into equally sized regions, and every region carves out a single category. For example for agent 3 the category category-4 covers the interval [0,0.5] on feature channel ball-y2 (Figure 6.6). The set of all categories of an

category	description	channel	bottom	top	score
category-8	moves backwards	delta-x	0.00	0.50	1.00
category-10	moves rightwards	delta-y	0.00	0.50	1.00
category-1	moves leftwards	roll-angle	0.50	1.00	1.00
category-2	moves rightwards	roll-angle	0.00	0.50	1.00
category-3	ends left	ball-y2	0.50	1.00	1.00
category-7	moves forwards	delta-x	0.50	1.00	1.00
category-4	ends right	ball-y2	0.00	0.50	0.86
category-9	moves leftwards	delta-y	0.50	1.00	0.85
category-22	gets closer	delta-d	0.00	0.50	0.73
category-6	ends behind	ball-x2	0.00	0.50	0.55
category-21	moves away	delta-d	0.50	1.00	0.55
category-16	ends right	ball-a2	0.00	0.50	0.50
category-19	starts far	ball-d1	0.50	1.00	0.51
category-14	moves short	roll-distance	0.00	0.50	0.50
category-17	starts in front	ball-x1	0.50	1.00	0.46
category-15	ends left	ball-a2	0.50	1.00	0.45
category-11	ends far	ball-d2	0.50	1.00	0.44
category-20	starts close	ball-d1	0.00	0.50	0.43
category-5	ends in front	ball-x2	0.50	1.00	0.42
category-18	starts behind	ball-x1	0.00	0.50	0.25
category-13	moves long	roll-distance	0.50	1.00	0.22
category-12	ends close	ball-d2	0.00	0.50	0.21

Fig. 6.6. The ontology of agent 3 after 4412 games

agent is called his ontology. Every category in the ontology has a score which is based on past success in the language games. Through adjustments of the score, agents progressively become aligned because the score also reflects not only the categories that are relevant in the scenes that they encounter but also those that are commonly used in the group.

In order to find a discriminating category, the categories for the most salient feature(s) are computed and then those categories are retained that are unique for the topic. In the present example, this is *the ball ends right* (category-4). When there is no discriminating category for the most salient feature channels in the ontology, the ontology is extended by refining a category applicable to the topic. Refinement of a category *c* happens by dividing the region of *c* into two equally sized subregions, which then yield two new subcategories. In the current experiment, the tree depth of the ontology never had to go deeper than one, however.

We use predicate-calculus notation (in prefix) to display the 'meaning' that is being expressed by the speaker (and reconstructed by the hearer). The predicates consist of all the categories in the ontology of the agent and the arguments are the event and the truth value. Here is an example: (category-4 event-16462 t).

6.1.3 Perspective reversal by the speaker

In some of the experiments we investigate the role of perspective alignment and perspective reversal. As mentioned earlier, we have endowed the agents with the capacity for egocentric perspective transformation, so that they can not only build up a situation model of themselves but also a model of what the other robot is supposed to see. If that is the case, the speaker can check whether the discriminating category of the scene which is valid for his own situation model also holds for that of the hearer. If so, perspective does not need to be marked (the perception of that feature of the scene is shared). Otherwise, the meaning to be expressed is extended with an additional predicate ('own-perspective') to specify that the perspective used is the speaker's. In the example, the category is not discriminative for the situation model from the hearer's perspective; in fact, it does not even hold in this model (the ball moves to the left in both events for the hearer). Hence the meaning is expanded by a perspective indicator:

(category-4 event-16462 t)
(own-perspective event-16462 t).

Alternatively, the speaker can completely conceptualize the scene from the viewpoint of the hearer and will choose it if it can be done with a more salient feature channel and based on a more established category. As Figure 6.3 shows (left bottom), for the assumed perspective of the hearer (robot B) the change in x position (channel delta-x) is the most salient channel, and the appropriate category (which happens to be category-7 or *moves forward*) can now be used. Meanings are ranked based on saliency and category score. The description with the highest score is then used in lexicalization. For the present case we have:

(category-4 event-16462 t) 0.393; from own perspective
(category-7 event-16462 t) 0.363; from other perspective.

So the first meaning is the best one from the viewpoint of conceptualization.

In the third experiment to be discussed later, the perspective is explicitly marked, which implies that it must be part of the meaning transmitted from the conceptualization subsystem to the lexical subsystem. Perspective is represented with two predicates, own-perspective and other-perspective, as in:

(category-7 event-16462 t)
(other-perspective event-16462 t).

6.1.4 The lexicon for the speaker

Each agent has a linguistic inventory, the lexicon (Figure 6.7). It is a bi-directional associative memory that associates abstract meanings (predicates and arguments with variables) with forms (words). Each association has a weight which acts as

score	form	meaning
1.00	*patide*	category-10
1.00	*kugizu*	category-8
1.00	*sotewu*	category-11
1.00	*remibu*	other-perspective
1.00	*lipome*	category-22
1.00	*livego*	category-1
1.00	*suvuko*	category-2
1.00	*bezura*	category-9
0.95	*lopapa*	category-3
0.95	*votozu*	own-perspective
0.85	*xapipu*	category-6
0.50	*fupowi*	category-4
0.30	*voxuna*	category-15
0.25	*naxopo*	category-16
0.20	*bikagi*	other-perspective category-8
0.15	*nodafo*	category-21

Fig. 6.7. The lexicon of agent 3 after 4412 games

a score reflecting how well the word involved had success in previous language games. We know from many earlier experiments that a reinforcement learning approach using lateral inhibition is an effective way to self-organize a lexicon (Steels, 2001). The speaker selects the smallest set of words that covers the complete meaning to be expressed (in the present example this is *fupowi votozu*). In case there are alternative solutions, the words with the highest score are used. Whenever the speaker does not have a word for the whole meaning or part of it, a new word is invented by combining random syllables and associating them with the uncovered meaning.

6.1.5 Lexicon lookup and conceptualization by the hearer

The hearer uses the same knowledge sources (lexicon and ontology) but in the reverse direction. He looks up the words in the lexicon and reconstructs the possible meanings. Usually there are several possibilities as words may be ambiguous. Next he attempts to interpret these meanings by matching them against the (reconstructed) situation model of the speaker and then his own situation model.

6.1.6 Feedback

A game is considered as successful if the hearer knows all the words in the utterance and if the extracted meanings are true and discriminating for the current

#	agent	meaning speaker	rough translation	utterance	rough translation	meaning hearer	agent	game status
5000	2	other-perspective category-3	from other persp. ends left	*gugita titelu*	from other persp. ends left	category-3 other-perspective	5	succeed
5001	4	other-perspective category-1	from other persp. moves leftwards	*gugita tenafa*	from other persp. moves leftwards	other-perspective category-1	5	succeed
5002	3	category-1	moves leftwards	*tenafa*	moves leftwards	category-1	4	succeed
5003	3	own-perspective category-1	from own persp. moves leftwards	*pugiza tenafa*	from own persp. moves leftwards	category-1 own-perspective	4	succeed
5004	3	own-perspective category-2	from own persp. moves rightwards	*pugiza ganagu*	?	own-perspective	1	fail
5005	4	other-perspective category-2	from other persp. moves rightwards	*gugita latawu*	from other persp. ends right	category-4 other-perspective	3	succeed
5006	4	other-perspective category-5	from other persp. ends in front	*gugita norolu*	?	other-perspective	3	fail
5007	4	own-perspective category-9	from own persp. moves leftwards	*pugiza mupofu*	from own persp. moves leftwards	own-perspective category-9	1	succeed
5008	2	other-perspective category-2	from other persp. moves rightwards	*gugita ganagu*	from other persp. ends right	category-4 other-perspective	4	succeed
5009	4	other-perspective category-8	from other persp. moves backwards	*sabesa gugita*	?	other-perspective	1	fail
5010	2	own-perspective category-2	from own persp. moves rightwards	*pugiza ganagu*	?	own-perspective	1	fail
5011	5	category-7	moves forwards	*ladole*	moves forwards	category-7	1	succeed
5012	1	own-perspective category-1	from own persp. moves leftwards	*pugiza tenafa*	from own persp. moves leftwards	category-1 own-perspective	2	succeed
5013	1	other-perspective category-7	from other persp. moves forwards	*ladole gugita*	from other persp. moves forwards	other-perspective category-7	4	succeed
5014	3	category-7	moves forwards	*ladole*	moves forwards	category-7	4	succeed
5015	5	category-4	ends right	*ganagu*	moves rightwards	category-2	3	succeed
5016	1	own-perspective category-8	from other persp. moves backwards	*kofunu pugiza*	?	own-perspective	2	fail
5017	2	own-perspective category-3	from own persp. ends left	*pugiza titelu*	from own persp. ends left	category-3 own-perspective	5	succeed
5018	5	other-perspective category-4	from other persp. ends right	*gugita ganagu*	from other persp. ends right	category-4 other-perspective	2	succeed
5019	4	other-perspective category-2	from other persp. moves rightwards	*gugita latawu*	from other persp. moves rightwards	category-2 other-perspective	5	succeed
5020	2	other-perspective category-1	from other persp. moves leftwards	*gugita tenafa*	from other persp. moves leftwards	category-1 other-perspective	1	succeed

Fig. 6.8. Subsequent interactions in a population of five agents (games 5000–5020)

event. Everything else is considered as a failure. Communicative success is the only measure that drives the coherence of perceptual categories and lexical items among the agents of a population. Therefore, each category and meaning–form association has a score that reflects its overall success in communication.

After a successful game, the score of the lexical entries that were used for production or parsing is increased by 0.05. At the same time, the scores of competing lexical entries with the same form but different meanings are decreased by 0.05 (lateral inhibition). In case of a failure, the score of the involved items is decreased by 0.05. This scoring adjustment not only acts as a reinforcement learning mechanism but also as a priming mechanism so that agents gradually align their lexicons in consecutive games.

When the hearer does not know one of the words of the utterance, he conceptualizes the scene himself by using the meanings that are already known from the utterance and the additional meanings are then associated with the unknown word(s). This step leads to a kind of replicator dynamics, because words invented or used by the speaker become part of the repertoire of the hearer who could then use it in subsequent interactions.

Agents not only play a single game but take turns playing games (see Figure 6.8), and it is through these consecutive games that a consensus gradually arises in the group. Not only do the lexicons become aligned, but also the ontologies. More and more agents will prefer to use the same conceptualization in the same sort of circumstance and use similar words for similar meanings.

6.2 Experimental Results for Perspective Alignment

As stated in the introduction, we want to show why perspective is relevant in spatial language and how agents manage to align and mark perspective.

6.2.1 The need to consider perspective

We begin with a first experiment to argue the first point stated in Section 1: *As soon as agents are embodied, they necessarily have a specific view on the world and spatial language becomes impossible without considering perspective.* It is straightforward to do a very clear experiment concerning this claim with the mechanisms introduced in the previous section.

First we show in a baseline condition that the cognitive mechanisms proposed earlier for behaviour, perception, conceptualization, and lexicalization are adequate when both agents engaged in a dialogue perceive the scene through the same camera and hence have exactly the same situation model. Although there is still some form of embodiment here (in the sense of using real vision and real-world action), it is not 'real' embodiment in the sense of each agent having his own body. As shown in Figure 6.9, communicative success quickly increases to 90% and the

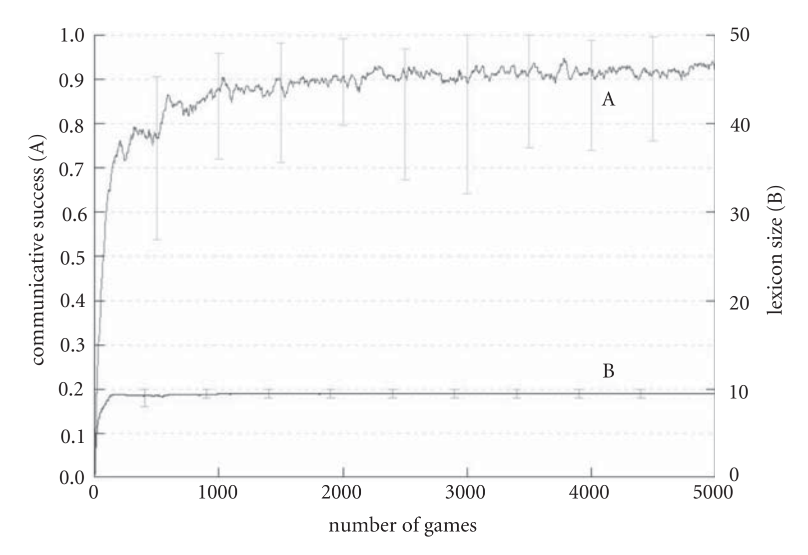


Fig. 6.9. Agents have the same sensory information and hence share their situation model. They quickly self-organize a lexicon and ontology.

average lexicon size of the agents is ten. These results are based on ten runs of 5000 language games each. We show the average and the variance. So this experiment shows clearly that the mechanisms proposed here work properly.

In the next condition, the agents perceive the scene through their own camera but they do not take perspective into account. The results are shown in Figure 6.10. Now they do not manage to agree on a shared set of spatial terms. Communicative success does not reach 10%. This clearly proves the first hypothesis, namely that grounded spatial language without perspective does not lead to the bootstrapping of a successful communication system for this kind of communicative task.

6.2.2 Perspective without marking

The next argument we put forward is the following: *Perspective alignment is possible when the agents are endowed with two abilities: (i) to see where the other one is located, and (ii) to perform a geometric transformation known as Egocentric Perspective Transform*. Both of these abilities have been implemented for the robots as explained earlier and so it is now possible to do an experiment that exercises these mechanisms.

When agents are able to perform egocentric perspective transformation and when the allocentric situation model is used as well in conceptualization, a successful communication system indeed emerges (Figure 6.11). Communicative

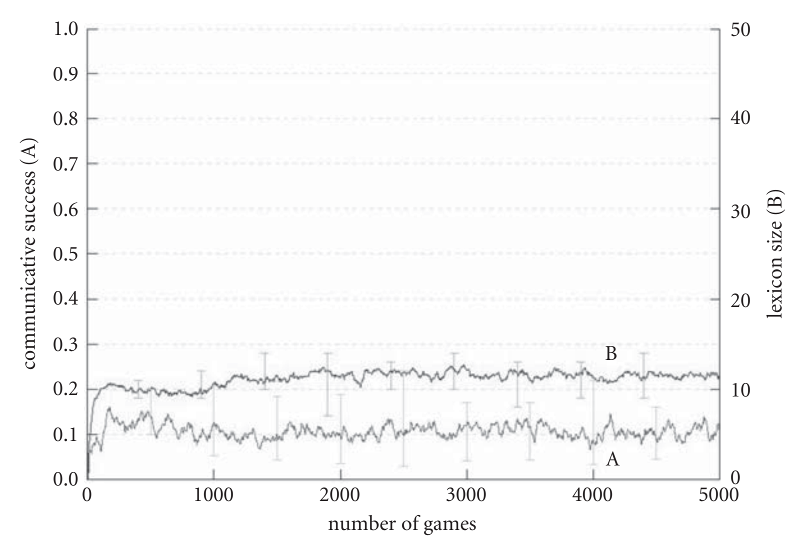


Fig. 6.10. Agents do not share sensory stimuli and do not consider perspective. The system does not come off the ground.

success again reaches 90% and the lexicon stabilizes. This is even without marking perspective. The reason the agents are nevertheless successful is because they continuously check from each perspective what a possible meaning or a possible interpretation might be. So we have an answer to the question of how it is possible for two partners in dialogue to align perspective even if there is no explicit marking.

6.2.3 The role of perspective marking

We now perform a third experiment to examine the third thesis: *Perspective alignment takes less cognitive effort if perspective is marked.* In the previous experiment, the hearer has to guess (by trying to interpret the utterance for both perspectives) which perspective was used, and the speaker has to compute both perspectives to make sure he chooses the one that will have most success with the hearer. This obviously results in a higher cognitive effort for the hearer. Cognitive effort is defined as the average number of additional perspective transformations that the hearer has to perform (cf. Figure 6.11).

Now we change the language architecture slightly for each agent. The chosen perspective is explicitly made a part of the meaning so that it becomes lexicalized; for example, as in:

(category-7 event-16462 t)

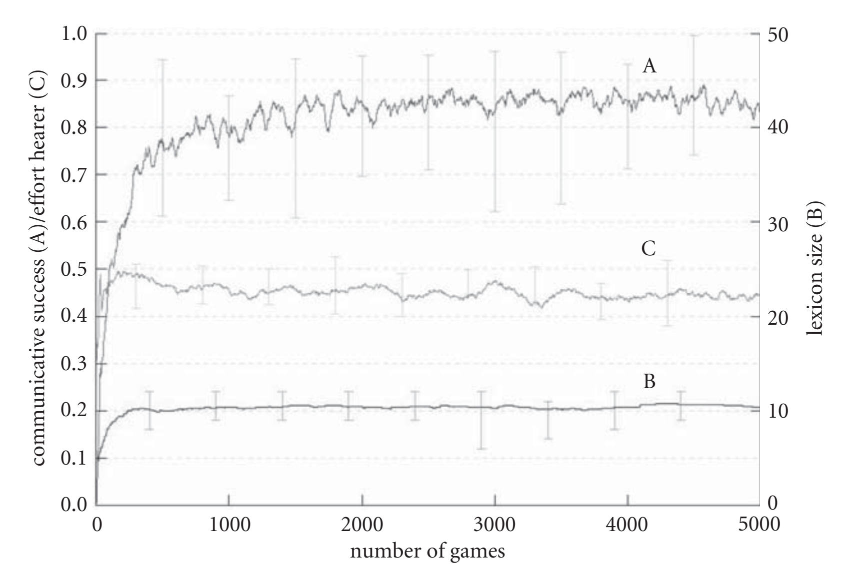


Fig. 6.11. Agents are able to adopt the interlocutor perspective but do not mark their perspective choice in language. They manage again to self-organize a spatial language system.

(other-perspective event-16462 t)

and this will automatically lead to an expression of perspective. Note that the lexicon formation process contains no bias towards the use of perspective markers. It tries to cover the complete meaning with the smallest number of words and invents new words for parts that are not yet covered. Nevertheless, we see that separate words emerge for perspective in addition to words where perspective is part of the lexicalization of the predicate. This is similar to natural language where in 'the ball to my left', 'my' is a general indicator of perspective, whereas in the German 'hinein' ('into' from outside perspective) versus 'herein' ('into' from inside perspective) or English 'come' and 'go', perspective is integrated in the individual word. So this experiment explains why perspective marking occurs in human languages and why sometimes we find specific words for it.

As shown in Figure 6.12, communicative success remains high but the cognitive effort dramatically decreases compared to the earlier experiment. Communicative success is not as high as in the previous experiment without perspective marking (Figure 6.11). This is due to the fact that the learning problem is harder for the agents as they additionally have to agree on a set of perspective markers or words that incorporate domain categories and a perspective marker, but if we look at a

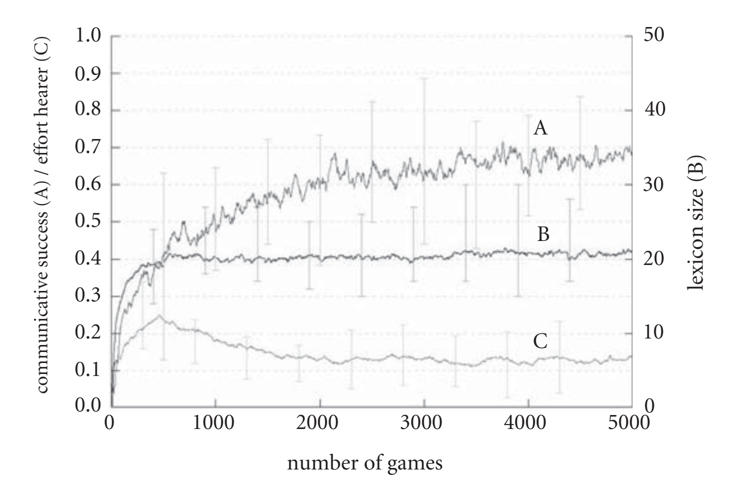


Fig. 6.12. Agents now additionally mark their perspective choice in language. This maintains communicative success but results in a decrease of cognitive effort.

longer series of games we see that a similar level of success is reached. Moreover, we have a more compact lexicon than in the previous experiment.

6.3 Conclusion

This chapter is significant from two points of view. On the one hand, it shows a novel way to investigate spatial language and perspective, namely by doing experiments in which physically embodied agents (robots) are endowed with a 'language faculty' that allows them to bootstrap a communication system autonomously (i.e. without human intervention) and from scratch. This rather new methodology is complementary to empirical observations of human dialogue and helps us to develop and test 'mechanistic' theories of dialogue (Cangelosi and Parisi, 2002). On the other hand, we could show very precisely why perspective is essential for spatial language, how speaker and hearer could align perspective—even without marking—and why and how perspective could become explicitly marked as part of spatial dialogue.

Acknowledgements

The experiments are based on the Fluid Construction Grammar framework (Steels and De Beule, 2006), which is highly complex software for language processing to which Nicolas Neubauer and Joachim De Beule made major

contributions. The authors also thank the members of the 'GermanTeam' for providing their robot soccer software and Remi van Trijp for editorial help with the chapter. This research was funded and carried out at the Sony Computer Science Laboratory in Paris with additional funding from the EU FET ECAgents Project IST1940. Comments may be sent to the authors at steels@arti.vub.ac.be.

7

Formulating Spatial Descriptions across Various Dialogue Contexts

LAURA A. CARLSON and PATRICK L. HILL

7.1 **Introduction**

Consider the office scene in Figure 7.1 depicting numerous objects, including a calculator, a book, a stapler, a ruler, and a bookshelf. For such a scene, if one asks 'Where is the calculator?', the following subset of responses is possible:

(1) The calculator is behind and to the right of the book.
(2) The calculator is behind the stapler.
(3) The calculator is somewhat in front of the bookshelf.
(4) The calculator is somewhat to the left of the ruler.

There are three components to each of these spatial descriptions: the *target* (the object being located, i.e. the calculator); the *reference object* (the object from which the target's location is described, i.e. the stapler, the book, the bookshelf, or the ruler); and the *spatial term*(*s*) that convey the relation between the target and reference object (i.e. behind, right, front, left). Utterances 1–4 constitute responses within a relation judgement task (Logan and Sadler, 1996) in which the goal of the speaker is to indicate to an addressee the location of a pre-specified target (i.e. the calculator). Thus, for this type of communicative task, it is up to the speaker to select an appropriate reference object and an appropriate spatial term that will make it easy for the addressee to find the target.

7.1.1 Three possibilities for selecting reference objects and spatial terms

The goal of the current chapter is to investigate how these components are selected across various dialogue contexts. There are three possibilities. First, a speaker could select the most salient reference object. This would then restrict the selection of a spatial term to those that convey the relation between the target and the selected reference object. This option is referred to as *Reference Object First*. This is exemplified by Utterance (1), in which the book is selected as the reference object. The book is perceptually prominent with respect to its size, shape, and brightly coloured cover, but its spatial relation with respect to the calculator is complex,

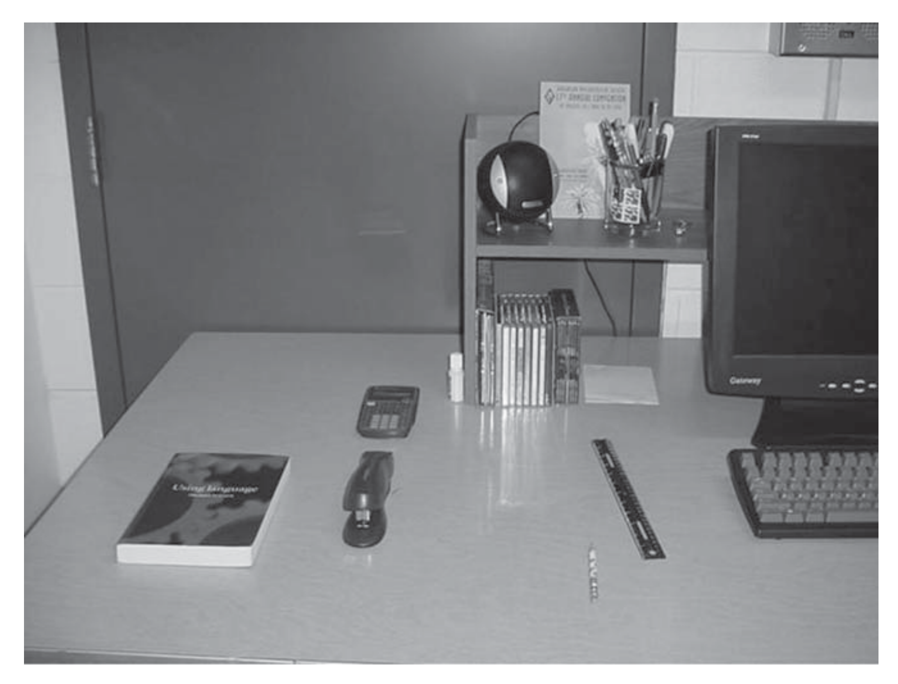


Fig. 7.1. Sample display of objects for which Utterances 1–4 indicate the location of the calculator.

and must be described with a combination of spatial terms. This option is consistent with a speaker's underlying assumption that the reference object should be known or easily found (Clark, 1996; Sperber and Wilson, 1995) by the addressee based on salient features including perceptual attributes (size, permanence of location, geometric complexity; see Talmy, 1983), conceptual features (an object recently mentioned, inanimacy; see de Vega, Rodrigo, Ato and Dehn, 2001), and semantic association with the target (functional relations; see Carlson-Radvansky and Tang, 2000). Research in referential communication similarly suggests that salient features that distinguish an object from its neighbours are selected for expression (Clark, 1996; Deutsch and Pechmann, 1982; Mangold and Pobel, 1988; Olson, 1970; Sperber and Wilson, 1995; Tenbrink, 2005).

Second, a speaker could select an object that is placed in a prototypical or 'good' spatial relation (Hayward and Tarr, 1995; Logan and Sadler, 1996) with respect to the target. This would then restrict the selection of a reference object to the subset of objects that are in this prototypical relation. This option is referred to as *Spatial Term First*. This idea is exemplified by Utterance (2), in which the spatial relation *behind* is selected because the target falls along an axis extending from a possible reference object (stapler). This selection of the spatial term *behind* restricts the speakers to using the stapler as a reference object, despite it not being salient

on perceptual, conceptual, or functional grounds. More generally, spatial terms parse space around reference objects into regions (Herskovits, 1986; Langacker, 1987; Miller and Johnson-Laird, 1976) that are graded, with some placements considered more acceptable than others (Hayward and Tarr, 1995; Logan and Sadler, 1996). Typically, for projective spatial terms like *front*, locations that fall along a principal axis (i.e. an axis centred on and extending out from an object's front or a viewer's front: e.g. Landau, 1996; Levinson, 1996) are considered 'good' or 'canonical', and can be described by selecting a single spatial term (Franklin, Henkel, and Zangas, 1995). In contrast, those falling to the side of the axis or close to the border of the region are considered 'acceptable' and are typically described by a combination of terms or by hedging (Franklin, Henkel, and Zangas, 1995; Lakoff, 1972). A preference for describing object locations using a single spatial term rather than a combination of terms and hedges is consistent with the Gricean maxims of informativeness and quantity (Grice, 1975). A prioritization of the spatial relations among the objects is also supported by work showing a preference for terms that differentiate the target's relation from other object relations (Tenbrink, 2005), and is consistent with the claim that specification of positional information may better identify an object than reference to its attributes, at least within human–robot interactions (Moratz, Tenbrink, Bateman, and Fischer, 2003).

Third, a speaker could jointly consider object/term pairs, selecting the combination that offers the best object in the best relation. This option is referred to as *Joint Selection*. This is exemplified in the contrast between Utterances (3) and (4), both of which contain a modified spatial term and a reference object. However, Utterance (3) would be preferred on this account for two reasons: first, the *front* relation is easier to compute than the *right* relation (e.g. Bryant, Tversky, and Franklin, 1992; Clark, 1973; Farrell, 1979; Franklin and Tversky, 1990; Logan, 1995, Maki, Grandy and Hauge, 1979). Second, the bookshelf as a reference object is larger and more permanently located (Talmy, 1983) than the ruler. Recent research in attention suggests that the processing of objects and locations may not be as separable as previously assumed (e.g. Logan, 1996; Soto and Blanco, 2004; Tsal and Lavie, 1988, 1993). Given the link between attention and spatial language (Langacker, 1987; Logan 1994, 1995; Regier and Carlson, 2001; Slobin, 1996; Talmy, 1996, 2000; for review, see Carlson and Logan, 2005), it is plausible that both object and term information would be jointly considered when formulating spatial descriptions. This option is also generally consistent with the idea that the selection of a spatial term to describe a given relation depends in part on the identity of the objects, their context, and their potential interaction (Carlson-Radvansky and Radvansky, 1996; Coventry, 1999; Coventry and Garrod, 2004). Finally, it should be noted that this option is a hybrid of the first two options, and is thus not mutually exclusive. That is, depending on the relative weights of object and term information, Joint Selection could make analogous predictions to Reference Object First (sole influence of object information and no influence of

term information), Spatial Term First (sole influence of term information and no influence of item information), or a hybrid with some influence of both sources of information.

7.1.2 Overview of studies conducted

As described in Sections 7.2–7.4, we conducted a series of studies to assess these possibilities, examining the production of responses to queries of the type 'Where is the target?' that were asked in conjunction with a depicted display or in the context of a real interaction. Across studies, we varied the dialogue context, including the presence of an addressee, the modality in which the participant responded, and the number of descriptions elicited per participant. The manipulation involving the presence of the addressee was motivated by the idea that monologue and dialogue contexts form a continuum (Pickering and Garrod, 2004), varying with respect to the availability of context and feedback. Given that speakers use common ground to facilitate apprehension for the listener (Clark, 1996; Isaacs and Clark, 1987; Schober, 1993; Sperber and Wilson, 1995), we wanted to assess whether the presence of a shared dialogue context would change the salience of objects and/or spatial relations. This is particularly important as traditional work in monologue contexts may not necessarily generalize to dialogue contexts (Pickering and Garrod, 2004). The manipulation of modality (oral, written, from memory) was based on research suggesting that different speech acts involve different goals, and these may differentially impact selection (Clark, 1996; Miller and Johnson-Laird, 1976; Sperber and Wilson, 1995). Finally, the manipulation of the number of descriptions that was elicited per participant was based on research suggesting that strategies for formulating descriptions may develop across trials (e.g. Brennan and Clark, 1996; Levelt, 1989; see also Vorwerg, this volume, for evidence of consistent use of spatial elements within discourse). Within each of these scenarios, the speaker and the present or presumed addressee faced the to-be-described scene from the same perspective. This is important because different viewpoints between speakers and addressees raise the additional complication of which perspective the speaker should adopt. Such offsets have been shown to significantly impact preferences for particular spatial terms (e.g. Schober, 1993).

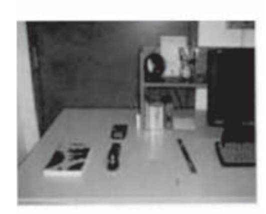
Within each study, we manipulated the location of objects to assess whether participants were consistently selecting the most salient object (consistent with Reference Object First), the object in the best spatial relation (consistent with Spatial Term First), or the best object/term combination (consistent with Joint Selection). For example, consider the displays shown in Figure 7.2. The displays vary in the placement of the target (the calculator) and placement of the candidate reference objects (book and stapler). Note that the book is larger than the other candidate objects placed on the desk; it is also uniquely coloured with a bright red pattern on the front cover. These features make it the perceptually salient object.

Where is the calculator?



Predicted reference object

Reference Object First:	Book
Spatial Term First:	Book
Joint Selection:	Book



Predicted reference object

Reference Object First:	Book
Spatial Term First:	Stapler
Joint Selection:	Book/Stapler



Predicted reference object

Reference Object First:	Book
Spatial Term First:	Book/Stapler
Joint Selection:	Book/Stapler



Predicted reference object

Reference Object First:	Book
Spatial Term First:	Book/Stapler
Joint Selection:	Book/Stapler

Fig. 7.2. Predictions from the Reference Object First, Spatial Term First, and Joint Selection hypotheses for the critical manipulations of object features and location for the studies presented in Sections 7.2–7.4.

In the first display at the top of Figure 7.2, the calculator is in a good *behind* relation with respect to the book. All three hypotheses predict that the book will be selected as the reference object. For Reference Object First, this is because the book is the perceptually salient object; for Spatial Term First, this is because the book stands in the best spatial relation with respect to the calculator; and for Joint Selection, this is because the book represents the best object/term combination.

In the second display in Figure 7.2, the calculator is in a good *behind* relation with respect to the stapler. Reference Object First predicts selection of the book because it is the salient object. Spatial Term First predicts selection of the stapler because it is in the best spatial relation. Joint Selection predicts a mixture—sometimes the book may be selected, and sometimes the stapler may be selected, with their proportions depending upon the relative weight of object and term information in the combination.

In the third and fourth displays in Figure 7.2, the calculator is placed in an acceptable *behind* relation with respect to both the book and the stapler. These displays allow us to assess whether object information will become important when there is no good spatial relation between the target and any candidate object. Reference Object First predicts selection of the book because it is the salient object. Spatial Term First predicts random selection of the book and stapler because neither is in a good spatial relation. Alternatively, there may be an increase in the use of other types of spatial descriptions. Joint Selection predicts a mixture; the extent to which the book might be selected more often than the stapler would illustrate an impact of object information.

Across the studies described in Sections 7.2–7.4, analogous sets of predictions can be generated from Object First, Spatial Term First, and Joint Selection hypotheses, although the particular configurations, the types of objects, and the dialogue contexts varied. More specifically, in Section 7.2 we describe production data elicited in response to displays consisting of isolated objects that varied in perceptual and conceptual salience and that were presented on a white background on a computer monitor. Each participant provided a large number (N = 128) of written descriptions, and there was no mention of an intended listener/reader. In Section 7.3 we describe production data elicited in response to a photograph of a real-world scene (as in Figure 7.1); there was an assumed but not co-present addressee (listener/reader), and participants provided only a single description that was either oral or written. In Section 7.4, we describe production data elicited in the course of a natural interaction in a real-time setting that mimicked the scene depicted in Figure 7.1 with a co-present and interactive listener.

7.2 Producing Multiple Written Spatial Descriptions of Targets in Multi-Object Arrays

Carlson and Hill (2008, Experiment 2) elicited spatial descriptions in response to queries of the form 'Where is the [target]' for computerized displays containing a varying number of isolated objects. Participants were provided with a sentence frame such as 'The [target] is ______________', and asked to type in their description on the keyboard. There was no mention of a prospective addressee, and no restrictions were placed on their responses in terms of length or content. Each participant completed 128 descriptions. Figure 7.3 shows a subset of three critical

displays using one of the eight object sets (hamburger, bottle of mustard, bottle of pesticide) that were used in the study. Each object set contained one object that was considered salient relative to the other objects on both perceptual and conceptual grounds; for the object set in Figure 7.3, the salient object was the hamburger. With respect to perceptual salience, this object was larger and of a different shape than the other two objects (mustard and pesticide); in contrast, these objects were matched in size and shape. Indeed, an independent group of participants verified that the salient object was perceptually distinct from the other objects. With respect to conceptual salience, this object participated in the same functional interaction with the other two objects as defined by object knowledge and dynamic-kinematic routines (Coventry and Garrod, 2004); for one of these objects, this action was typical (mustard and hamburger) and, for the other object, this action was atypical (pesticide and hamburger). For the displays in Figure 7.3, the target was the mustard; our main interest was whether on a given trial participants would select the salient object (e.g. the hamburger) or a nonsalient object (e.g. the pesticide) as the reference object.

Considering displays that contained both the salient and nonsalient objects (N = 4700), we coded each utterance for each participant as to whether it contained a reference object, yielding a subset of 2964 utterances. This set of utterances included a variety of types (see Carlson and Hill, 2008, for details), but we will focus on the reference object that was selected. The remaining utterances used environmental references, such as 'The mustard is in the centre'. Note that categorizing the whole set of utterances into different types is of interest generally, as it establishes a baseline for the frequency with which spatial descriptions including a reference object are used. Indeed, such variety is consistent with Tenbrink's (this volume) argument about the considerable flexibility and creativity with which speakers produce spatial descriptions of complex scenes. This issue will be further addressed in the discussion in Section 7.5. For now, we focus on descriptions that contain a reference object because these enable us to test among the Reference Object First, Spatial Term First, and Joint Selection hypotheses.

For the subset of descriptions that contained at least one reference object, we coded how often the salient object (e.g. hamburger) was selected as the reference object; the corresponding percentages are shown in Figure 7.3. For descriptions that contained more than one reference object, we coded the first object that was mentioned consistent with previous arguments that the first object is deemed primary by the speaker (Mainwaring, Tversky, Ohgishi, and Schiano, 2003; Taylor and Tversky, 1996).

In Figure 7.3, Panel A, the target mustard is in a good *above* relation with respect to the salient object (hamburger) and is in a good *right* relation with respect to the pesticide. The hamburger is picked as reference object on the vast majority of the trials. This preference could reflect selection of a salient reference object (hamburger) or selection of the good *above* relation over a good *right* relation consistent with *above* being the easier term to compute (e.g. Clark, 1973; Franklin

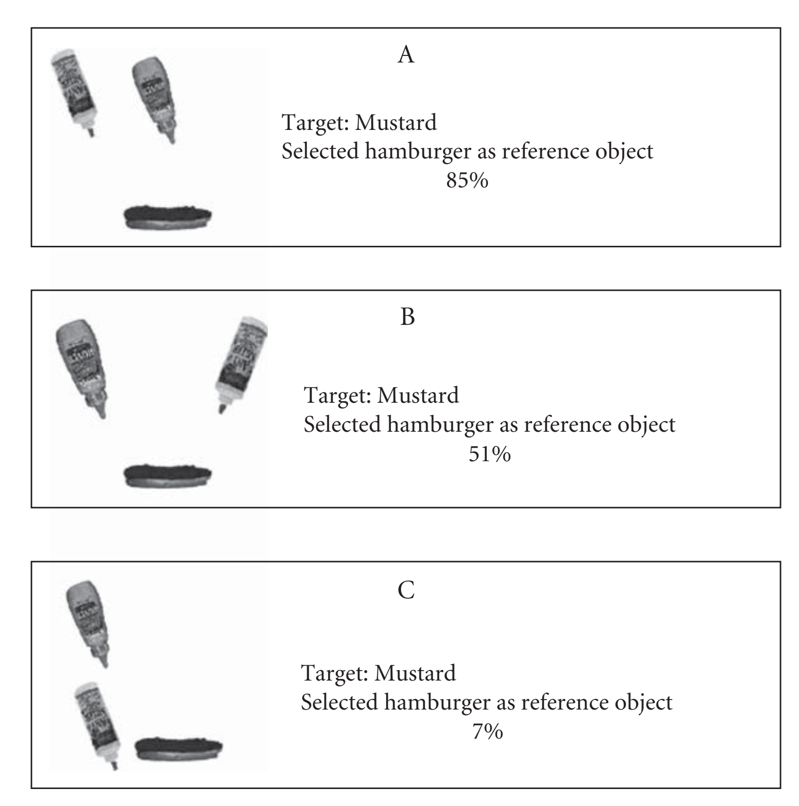


FIG. 7.3. Percentage of utterances in which the salient object (e.g. hamburger) was selected as the reference object, given descriptions that contained a reference object.

and Tversky, 1990; Logan, 1995). In Panel B, the target is in an acceptable *above* relation with respect to the salient object, and in a good *left* relation with respect to the nonsalient object. The hamburger is selected about half of the time, with the pesticide selected on the other half of the trials. The fact that both objects were used about equally often suggests that being in a good relation isn't always sufficient, particularly if the term is less preferred. In addition, these data are inconsistent with the Reference Object First hypothesis that would predict selection of the hamburger regardless of its location. Finally, in Panel C, the target is in a good *above* relation with respect to the pesticide and an acceptable *above* relation with respect to the hamburger. The hamburger is very seldom selected, indicating that when the term is held constant (both are 'above' relations), good relations are preferred to acceptable relations, regardless of the salience of the objects. On the whole, these data support a strong preference for selecting reference objects based on their spatial relations consistent with either the Spatial Term First or the Joint Selection hypotheses.

7.3 Producing a Spatial Description of a Target in a Photographed Scene

We further tested among these hypotheses by examining the production of spatial descriptions of a target (i.e. the calculator) within the photographed scenes shown in Figure 7.2 and reproduced in Figure 7.4. Each participant produced a single description, either written (N = 64) or oral (N = 36). Participants who wrote a description were shown the display and given a context that established a goal and reader: '*Mary needs to borrow Aaron's calculator. Use this picture of Aaron's room to find the calculator. Now tell Mary where to find it using the frame: "The calculator is* ____"'. Participants who orally produced the description were shown the display and told '*Mary needs to borrow a calculator. Tell Mary where it is*'. Although these contexts differ in length, both serve to establish a goal for the utterance and to identify a specific addressee. For completeness, the data are shown in Figure 7.4, reflecting the percentage of times a given object (book, stapler, desk, or other object) was selected as the reference object in their descriptions. Note, however, that we are particularly interested in the frequency of selecting the book (the salient object) or the stapler (the nonsalient object) as reference objects; these are the objects whose locations we manipulated across displays. We will return to the use of other objects within the descriptions in the discussion in Section 7.5.

In Figure 7.4, Panel A, the book (a good *front* relation with the calculator) was selected most frequently, in both written and oral descriptions. This could be due to its being the salient object based on its colour and size or its being in a good relation with respect to the calculator. In Panel B participants preferred the stapler (a good *front* relation with a less salient object) to the book (the salient object in an acceptable *front* relation). This rules out the Reference Object First hypothesis. In Panels C and D, the book and stapler are in acceptable relations. The book tends to be preferred in Panel C, and the stapler tends to be preferred in Panel D. These preferences can be described as a bias to mention the left-hand object first, consistent with Chan and Bergen's (2005) claim that reading direction has an influence on other cognitive processes. Note that the left-hand object is also the one located towards the edge of the desk, and thus may be salient by virtue of its location. Indeed, Gorniak and Roy (2004) argue for a preference for such spatially extreme objects. As with the study in Section 7.2, these data support a priority for selecting the spatial relation, and are consistent with either the Joint Selection hypothesis or the Spatial Term First hypothesis.

7.4 Producing a Spatial Description of a Target in a Naturalistic Interactive Dialogue Context

We further tested among the Reference Object First, Spatial Term First, and Joint Selection hypotheses by eliciting a single spatial description in a naturalistic

Where is the calculator?

A

Percentage selected as reference object

	Oral	Written
Book	56	50
Stapler	0	0
Desk	44	44
Other	0	6

B

Percentage selected as reference object

	Oral	Written
Book	0	0
Stapler	44	56
Desk	44	38
Other	12	6

C

Percentage selected as reference object

	Oral	Written
Book	44	13
Stapler	0	6
Desk	56	69
Other	0	12

D

Percentage selected as reference object

	Oral	Written
Book	11	6
Stapler	11	31
Desk	33	50
Other	45	13

Fig. 7.4. Percentage of utterances in which various objects (book, stapler, desk, other) were selected as reference objects, broken down as a function of oral or written production. These percentages were computed across descriptions that contained a reference object.

interactive dialogue context. The target was a set of lab keys. Participants entered the lab and were instructed to sit at a particular table. The experimenter used the keys to open a door in the lab to retrieve papers, thus drawing attention to the keys and identifying them as belonging to the experimenter. The experimenter then

handed the participant the papers, and placed the keys on a table on which there was also a red binder (the salient object that was akin to the book in the study from Section 7.3) and a stapler (the nonsalient object). The locations of the binder and the stapler were manipulated across trials to create the four critical configurations that correspond to those used in the study in Section 7.3 (see Figure 7.5). The experimenter walked to the other end of the lab, approached a locked door, tried to open it, and stated 'Oh, I lost my keys'. After an initial search for the keys, the experimenter asked the participant 'Where are my keys?' This was the critical query, and we were interested in the form of the response made by participants, with a particular focus on spatial descriptions of the keys that included a reference object. When necessary, the experimenter provided a further prompt, such as 'Where were they?' when the initial query did not elicit a spatial description (for example, the participant simply pointed at the keys and said 'Here').

In addition, we manipulated the conceptual salience of the objects by drawing attention to selected objects prior to the query about the keys. In the *binder-emphasis* context, the experimenter picked up the binder, retrieved a piece of paper, and then replaced it. In the *stapler-emphasis* context, the experimenter picked up the stapler, stapled the pages in his/her hand, and then replaced it. Drawing attention to these objects was intended to establish them in the common ground, enabling us to assess whether salience with respect to the discourse setting would influence reference object selection. There was also a *neutral* context that did not involve an interaction with one of the critical objects. Finally, after eliciting a spatial description, we moved participants into a separate room, and asked them to provide a written description of the location of the keys from memory.

Figure 7.5 presents data for 138 participants in the form of the percentage of spatial descriptions that included the binder, the stapler, other objects in the room, reference to themselves (labelled 'ego', for example 'in front of me') or some use of ego (labelled 'some ego', for example, 'by the door that I came through'), coded on the basis of the object that was first mentioned. These are broken down as a function of their spontaneous oral response and as a function of their response from memory. For some conditions there are too few utterances in the design for analysis. Thus, we will simply describe some general trends in the data.

In general, for the oral descriptions, note first that the percentages are quite small. Most participants responded by pointing and using an ego-based description or including a noncritical object such as the table. For this reason, the numbers within a column do not necessarily sum to 100%. This is consistent with Bangerter's (2004) findings that pointing is often used to jointly focus attention during an interaction, with such use replacing explicit reference to target location. We will return to the use of different types of spatial descriptions in the discussion of Section 7.5.

Second, note that the percentages are larger for the memory responses overall, with a notable increase in the inclusion of the desk in their descriptions. Third, for the purposes of the testing among the hypotheses, the critical data are the selection

A

	NEUTRAL (N = 21)		STAPLER (N = 10)	
	Oral	Memory	Oral	Memory
Binder	0	14	20	10
Stapler	0	0	0	0
Desk	24	62	30	50
Other	0	5	0	10
Ego	43	0	30	10
Some ego	0	0	0	0

B



	NEUTRAL (N = 10)		BINDER (N = 16)	
	Oral	Memory	Oral	Memory
Binder	0	20	6	13
Stapler	0	10	6	6
Desk	0	60	13	63
Other	0	0	0	0
Ego	50	0	50	6
Some ego	10	0	0	0

C



	NEUTRAL (N = 10)		BINDER (N = 19)	
	Oral	Memory	Oral	Memory
Binder	10	0	0	11
Stapler	0	0	0	0
Desk	20	90	37	79
Other	0	0	0	5
Ego	60	10	42	5
Some ego	0	0	0	0

D



	NEUTRAL (N = 11)		STAPLER (N = 30)	
	Oral	Memory	Oral	Memory
Binder	0	0	0	0
Stapler	0	9	0	0
Desk	9	82	20	86
Other	0	0	0	7
Ego	64	0	43	7
Some ego	0	0	0	0

Fig. 7.5. Percentage of utterances in which various objects (binder, stapler, book, other, ego) were selected as reference objects, broken down as a function of context (neutral, emphasize stapler, emphasize binder) and whether it was an oral description or a description written from memory. Percentages were computed across all types of descriptions.

of the binder (the salient object) and the stapler (the nonsalient object) as a function of context (neutral, binder-emphasis and stapler-emphasis) across the four configurations. In these data, patterns emerge that are consistent with the data from Sections 7.2 and 7.3. First, in Figure 7.5, Panel A, the binder (in a good *front* relation with the keys) is selected in both the neutral and the stapler-emphasis context. This indicates a preference to select a perceptually salient reference object and/or an object in a good relation. In Panel B, there is a tendency to sometimes select the stapler (a good *front* relation) (absent in Panel A). This tendency is inconsistent with the Reference Object First hypothesis. More generally, across Figure 7.5, Panels A and B, conceptually emphasizing an object seems to have no influence on its selection; rather, when an object stands in a better spatial relation, that object tends to be selected as a reference object. In Panel C, there is a preference for the binder in both the neutral and binder-emphasis contexts. This may be due to it being a salient object by virtue of its perceptual features or by virtue of its placement on the left side of the desk. Finally, in Panel D, there is a very slight preference for the stapler, the object on the left side of the display. Again, selection does not seem to be influenced by the conceptual salience manipulation. Although there are small percentages in some of these cells, the patterns of data are largely consistent with those from Sections 7.2 and 7.3, indicating a preference for objects in good locations. When there are no good relations in the display, object attributes (such as location in the scene but not conceptual salience) seem to play a role in determining selection. These data are thus consistent with the Spatial Term First Hypothesis or a version of the Joint Selection hypothesis in which spatial relational information is weighted more heavily than object features. It is interesting to note, in this context, that when there are no good relations among the primary objects (book, stapler, calculator) (as in Figure 7.5, Panels C and D), participants opt most strongly for spatial descriptions that include the desk or another object rather than use descriptions with these objects. Thus, object features alone seem to have little effect on the formulation of the spatial descriptions.

7.5 Discussion

Spatial descriptions were elicited in a variety of dialogue contexts in order to assess how speakers select reference objects and spatial terms. Across the series of experiments, we found that speakers favour object/term pairs for which the spatial relation between the target and the reference object is easily conveyed by a simple spatial term; when no such simple term is available, and holding constant the spatial relation among objects, speakers may have a weak preference to select a reference object based on its features. In contrast, conceptual salience (as manipulated here as an interaction with the object) does not seem to impact reference

object selection; note, however, that this conclusion is somewhat provisional due to the small percentages of utterances in some of the cells of the design.

On the whole, the data support the Spatial Term First hypothesis, in which the primary consideration during formulation is selection of a reference object in a good spatial relation. One limitation to this hypothesis is that it does not make a prediction as to which object should be selected when there are no good relations in the display. However, consistently, for the displays in Panels C and D in the various studies when selection could not be based on the spatial relations, there was the suggestion of a pattern based on object attributes (see also Carlson and Hill, 2008). As such, the data may be more compatible with a version of Joint Selection in which spatial relational information has a primary role, with object features having only a weak influence when there are candidate reference objects in good relations, but a stronger influence when there are no good relations.

The strong influence of the spatial relation between the target and the reference object is an interesting finding, because theoretical discussions often mention perceptual features of the objects as factors that may impact selection (Talmy, 1983). Thus, one contribution of this work is to suggest that the most important feature is the spatial relation that the object forms with the target. Note that this pattern was consistent across dialogue contexts, varied sets of stimuli, the assumed or actual presence of an addressee, and was independent of the number of spatial descriptions produced (128 versus 1). Thus, the pattern cannot be an artefact due to setting, the particular materials, the degree of interaction between speaker and addressee, or a strategy emerging across trials. That is, when spatial descriptions that include a reference object are examined across these settings, materials, and degree of interaction, a consistent preference for a reference object in a good spatial relation is observed.

Note that this does not necessarily mean that this is the preferred type of spatial description. Indeed, in the study described in Section 7.2, Carlson and Hill (2008) found that the most frequent description included a reference to the environment, not another object in the display. More generally, across the manipulations of setting, materials, and degree of interaction in these studies the preference for the use of descriptions with reference objects varied. It is an interesting but separate question to determine which sets of conditions are more or less likely to elicit which types of descriptions. For example, Tenbrink (this volume) documents considerable flexibility in the use of different types of spatial description in response to 'Which object' queries that contrast with the 'Where is' queries in the current work; in Tenbrink's results preferences for types of descriptions were found to vary as a function of display configuration, speaker, and addressee location, and distinctiveness of object shape.

The question under investigation in the current chapter is slightly different—it addresses how selection occurs for a given type of spatial description. A strength of the current work is that it addressed this question with the implementation of similar manipulations of object features and placements across contexts that

varied from more artificial (typing a series of spatial descriptions of computerized displays) to more naturalistic (spontaneously providing a spatial description in a real-world context). The consistency in the pattern of results across such settings suggests that the preference for selecting a reference object based on its spatial relations (when those are prototypical) is quite robust.

Despite the variability in setting, there are several open issues remaining. Future research needs to more formally characterize perceptual salience, and examine different dimensions of salience. For example, with a stronger manipulation of salience, it may be possible to observe a stronger influence of object features, particularly in the absence of good spatial relations. It would be surprising if there exists a strong enough manipulation of object features that would override the spatial relation, although this is worthy of assessment. In addition, the shared perspective of the speaker and addressee should be manipulated. It may be the case that when the perspective is shared, the spatial relational information has a priority; however, under conditions in which the interpretation of spatial terms is more ambiguous (such as when speaker and addressee perspective are offset), object features may play a stronger role. Finally, Schober (this volume) presents an investigation of the impact of differences in spatial ability on spatial perspective taking and so it may also be the case that object features may be preferred when the speakers and addressees differ in spatial ability. This will need further investigation.

8

Identifying Objects in English and German: a Contrastive Linguistic Analysis of Spatial Reference[1]

THORA TENBRINK

8.1 **Introduction**

The semantics and applicability of spatial projective terms such as *left, right, front, back, above*, and *below* have been of interest for researchers in the field of spatial cognition for several decades. These linguistic items provide insights into speakers' conceptualizations of spatial settings and their preferences in describing them linguistically. Their application involves a range of concepts, for instance, perspective choice (Steels and Loetzsch, this volume), reference systems (Levinson, 2003; Watson *et al.*, this volume) and spatial templates (i.e. applicability areas; Logan and Sadler, 1996; Carlson-Radvansky and Logan, 1997), and is influenced in a number of ways by contextual factors such as the functions or salience of the objects involved (Carlson and Hill, this volume; Carlson and van der Zee, 2005; Coventry and Garrod, 2004) as well as speakers' discourse strategies (Herrmann and Grabowski, 1994). A systematic overview of such findings is presented in Tenbrink (2007). The focus of this chapter is to examine the impact of changes in the spatial configuration on the linguistic choices that English and German speakers prefer or disprefer. While most earlier studies in spatial language choose control over ecological validity, the present approach uses a more naturalistic scenario allowing speakers to react spontaneously to a given spatial task. Schober (this volume) uses a similar approach in that he lets participants choose their own words to describe the spatial location of objects, and Carlson and Hill (this volume) discuss a range of scenarios varying in naturalness.

A major part of earlier research on projective terms has focused on one specific, central kind of discourse task, namely, one in which the position of one known object in relation to another is to be described, as in answering a question such as, 'Where is the object?'. Tenbrink (2005) presented the results of a web-based study in which a different kind of discourse task is explored with English native speakers,

[1] Funding from the Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged.

namely, that of referential identification as in answering the question, 'Which object do you mean?'. This kind of discourse task is prominent in human–robot interaction scenarios, including those targeted in our research project (SFB/TR 8 on Spatial Cognition; Shi and Tenbrink, this volume). In these scenarios, an object needs to be identified out of a range of competing objects. The level of granularity chosen for such descriptions depends on the presence and placement of the other candidates. Previous experiments also involving referential identification have been reported, for example, in Schober (1993) and Mainwaring *et al.* (2003), focusing mainly on perspective choice. Mainwaring *et al.*'s results furthermore highlight the impact of differences in the discourse task. These systematically influence speakers' choices of perspective, their usage of distance terms rather than projective expressions, and the degree of redundancy in their utterances. One recent study in which linguistic choices in spatial referential identification are addressed in considerable detail can be found in Gorniak and Roy (2004). Results show that a very frequent strategy is 'to refer to spatial extremes within groups of objects and to spatial regions in the scene' (p. 439). This is done, for example, by terms of distance such as *closest*, or by projective terms such as *in the front* or *on the left side*. As indicated by the authors' interpretation of such terms as 'spatial extremes', such expressions, though linguistically unmodified, refer to the object that is situated at the most extreme position as compared to other objects that may also be situated in the same spatial region, for example, within the left half of the picture. Another possibility is to refer to the extreme position using a projective superlative such as *leftmost*. Altogether, the results show the considerable degree of flexibility and creativity available to speakers in their spontaneous spatial descriptions in complex scenarios.

The present study presents a comparison of the results reported in Tenbrink (2005) for English with results gathered for the German language. I will describe to what extent German speakers adhere to similar discourse strategies to English speakers, and analyse in detail in what ways the linguistic representations diverge. As will be shown, a number of systematic differences can be accounted for by differences in language structure. However, general principles of spatial referential identification can be identified in both languages; these are in accord with earlier findings on contrastive reference (e.g. Herrmann and Deutsch, 1976).

Language-specific differences in spatial reference have previously been addressed most systematically by the Max Planck Institute for Psycholinguistics (MPIP) in Nijmegen (Levinson, 2003), which aims to examine possible correlations between language and cognition based on detailed comparative research in many different cultures, using 'Where' tasks and focusing on conceptual aspects. Further authors comparing English and German spatial reference (e.g. Carroll, 1997 and Weiß *et al.*, 1996) also investigate differences in the choice of reference frames. Herskovits (1986) (for English), Wunderlich and Herweg (1991), and Eschenbach (2005) (for German) present overviews of the syntactic repertory of

projective terms, together with detailed semantic analyses and further suggestions concerning applicability. The present approach, in contrast, addresses a detailed contrastive linguistic analysis of speakers' spontaneous choices in relation to a range of simple spatial configurations, using a 'Which' (referential identification) task.

8.2 Empirical Study: Method

The present account incorporates results of a large web-based empirical study (see Tenbrink, 2007), parts of which were previously reported in Tenbrink (2005). The study was made accessible at the 'Portal for Psychological Experiments on Language', maintained by Frank Keller, at www.language-experiments.org for two periods of time: the English version was online between 23 Sept. and 31 Dec. 2003; after that, the German version was online until 4 Oct. 2004. Participation was on a voluntary basis. Age effects were not tested for; the participants were predominantly between 15 and 50 years old. Several hundred participants from all over the world took part in this study.

Each of the participants completed 15 different randomly assigned tasks in randomized order out of a pool of 29 possible tasks which cover a range of different scenarios. The decision to limit the number of tasks for each participant to 15 was taken in order to minimize the time and effort required for participation. 28 of the 29 tasks consisted of a picture and an associated question triggering referential identification as discussed in the previous section. They belong to three conditions explained below that differ with respect to the possible perspectives on the configurations. Each participant received four questions in each of conditions 1 and 2, and six questions in condition 3. The only constraint with respect to the participants' contributions was that they were asked not to use counting in order to avoid collecting a large corpus of linguistic utterances that rely solely on counting rather than projective terms. Randomization of task assignment and order applies only inside conditions. The final task (a route instruction task) is not analysed here. After finishing the tasks, all participants were asked to answer some demographic questions.

In each condition, the same set of basic configurations of elements was shown (see Figures 8.1–8.4 below); the choice of configurations is explained in Tenbrink (2005). The configurations S6 and S7 (S for situation) were only added to the study at a later time. Therefore, there were no contributions by native English speakers and the number of contributions by German speakers was much lower for these configurations. They were added in order to gain information about the vertical and frontal axes, since most of the other pictures produced a bias towards the employment of the lateral axis for reference. Altogether, the seven basic configurations offered a range of spatial relationships that could be conceptualized and referred to in various ways. The variability offered by the different configurations

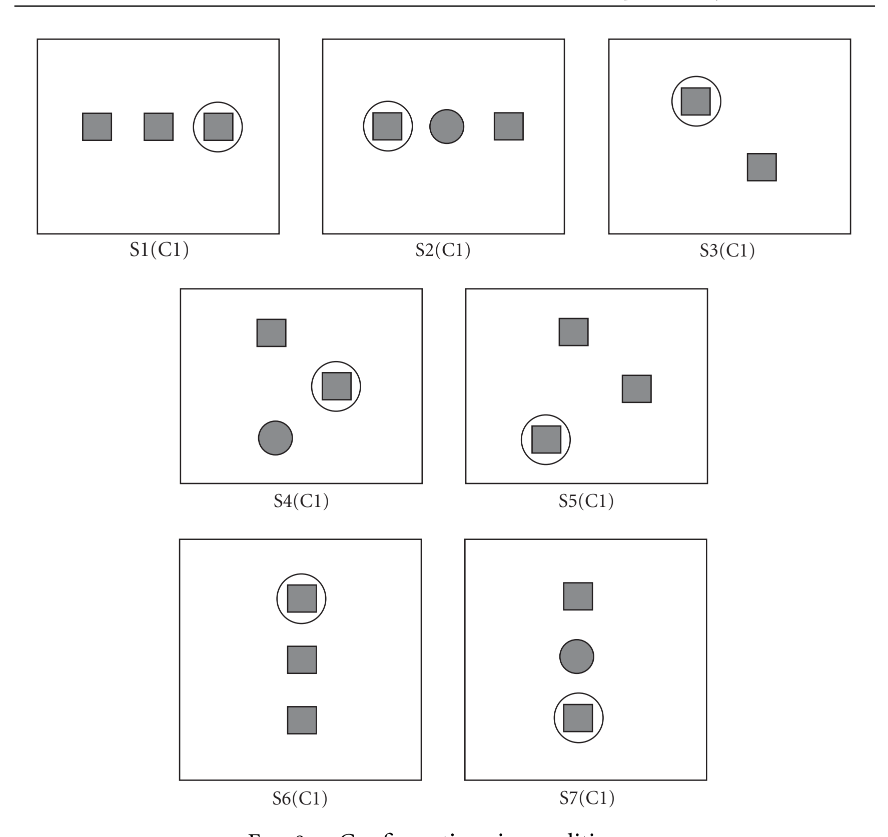


Fig. 8.1. Configurations in condition 1.

was further enhanced by the options for perspective, which were varied by the three conditions as follows.

In the first condition, participants were presented with pictures that only contained squares and circular elements (Figure 8.1). In each picture, one of the elements was marked by a circle. The question to be answered by the participants was simply (in the English version), 'Which element of the picture is marked with a circle?' In the second condition, an X appeared in the picture in addition to the elements (Figure 8.2). The English instruction was: 'Now imagine that you are looking at the figures from the position marked X. How do you describe now which element is marked with a circle?' The third condition was designed to simulate a real world setting as much as possible in order to enable a comparison to a human–robot interaction setting addressed in our project (e.g. Moratz and Tenbrink, 2006), where basically the same set of configurations was used. Here, the position of an interaction partner, Y, was added to the pictures. Additionally, both X and Y were assigned a view direction. For each of the scenarios, there were two

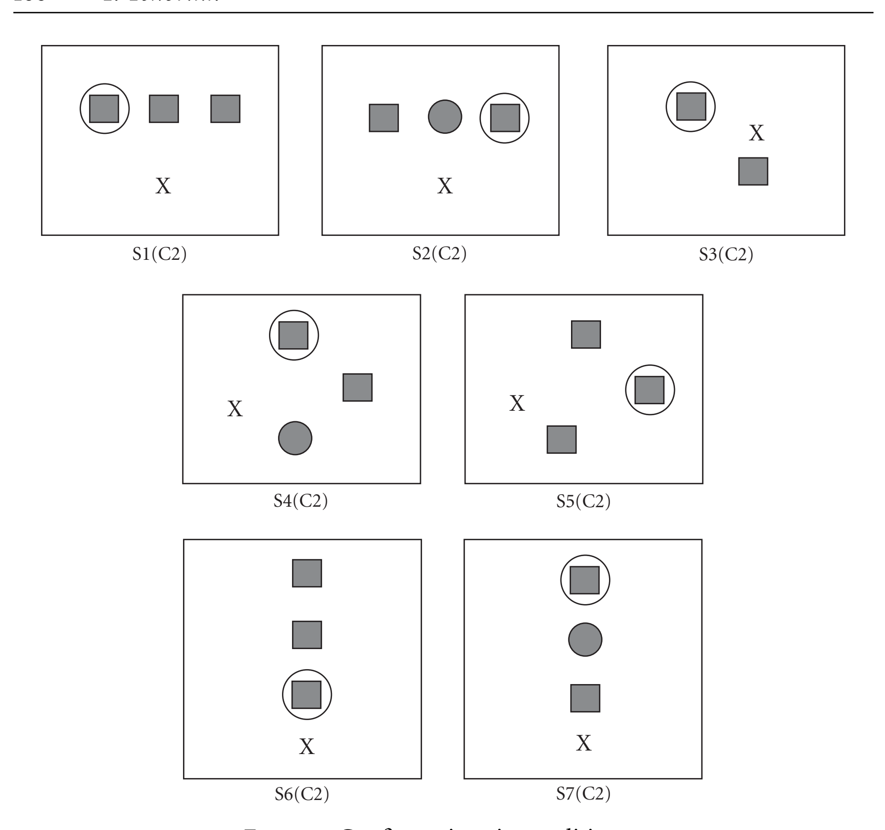


FIG. 8.2. Configurations in condition 2.

possibilities for the positions of X and Y (cf. condition 3A, Figure 8.3, vs. condition 3B, Figure 8.4) so that the number of configurations was twice the number in condition 3 compared to conditions 1 and 2. In each case, the participants read: 'Finally, please imagine that the figures are real-world objects. You are located at X, and now your task is to instruct person Y to go to the object marked with a circle. A star * shows the direction each of you is facing in.'

Thus, in this condition, view directions were given explicitly, and the participants were asked to imagine a dialogue situation. However, since there was no real interaction and no feedback from the interaction partner, grounding and alignment processes such as those described in Clark (1996) and Pickering and Garrod (2004) were ruled out, in a similar way to imagined-partner experimentation as reported in Herrmann and Grabowski (1994). The instructional task in condition 3 differed slightly from the previous ones. This situation required a lot of imagination by the participants; therefore, tasks in condition 3 were presented only after the first two conditions.

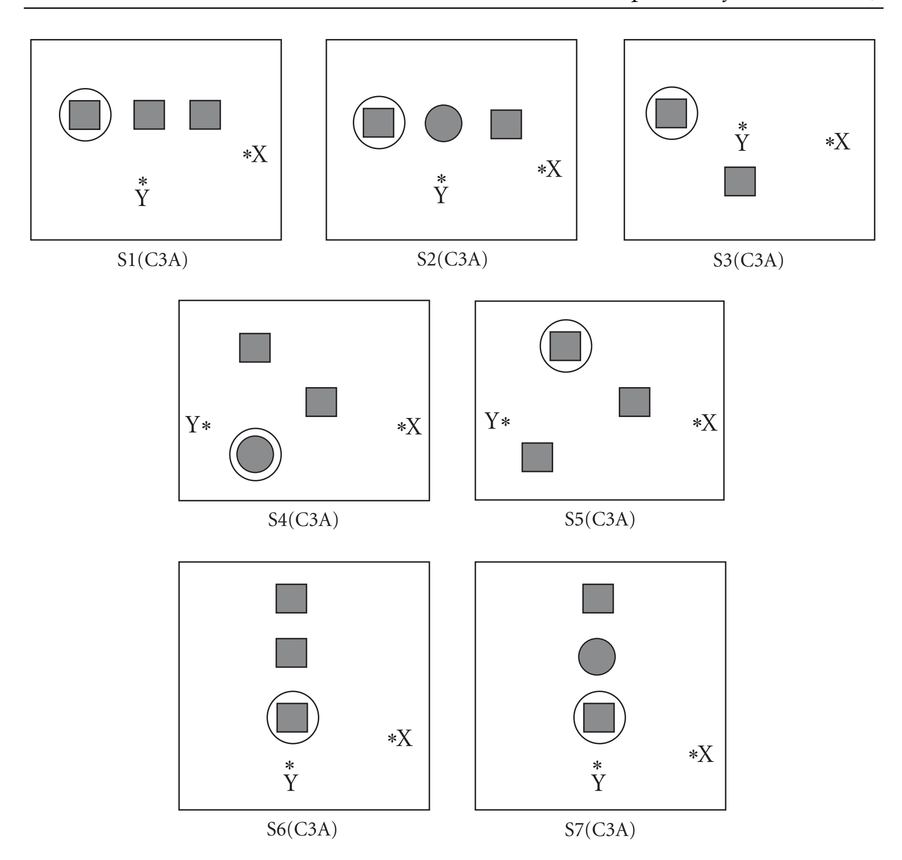


Fig. 8.3. Configurations in condition 3A.

A total of 2,332 German utterances produced by approximately 180 native speakers of German, plus 1,480 utterances produced by approximately 200 native English speakers were analysed. The contributions were extracted from the collected pool of data, annotated, and analysed linguistically. For each situation, the preferred linguistic options were identified and analysed in detail (see Tenbrink, 2005, for more details on motivation and research questions). The basic question underlying the analysis of each configuration is: *What linguistic options do speakers of the two languages prefer?*

Differences between situations were examined by comparing frequencies of linguistic categories, and explained on the basis of features of the configurations. Hypotheses generated in this way were tested further by comparing groups of situations sharing critical features, and examining the relevant features of the linguistic contributions. This procedure is regarded as a suitable and necessary preparation for ensuing experiments (using, for example, psycholinguistic methods), testing the hypotheses gained directly by varying configurations in relevant

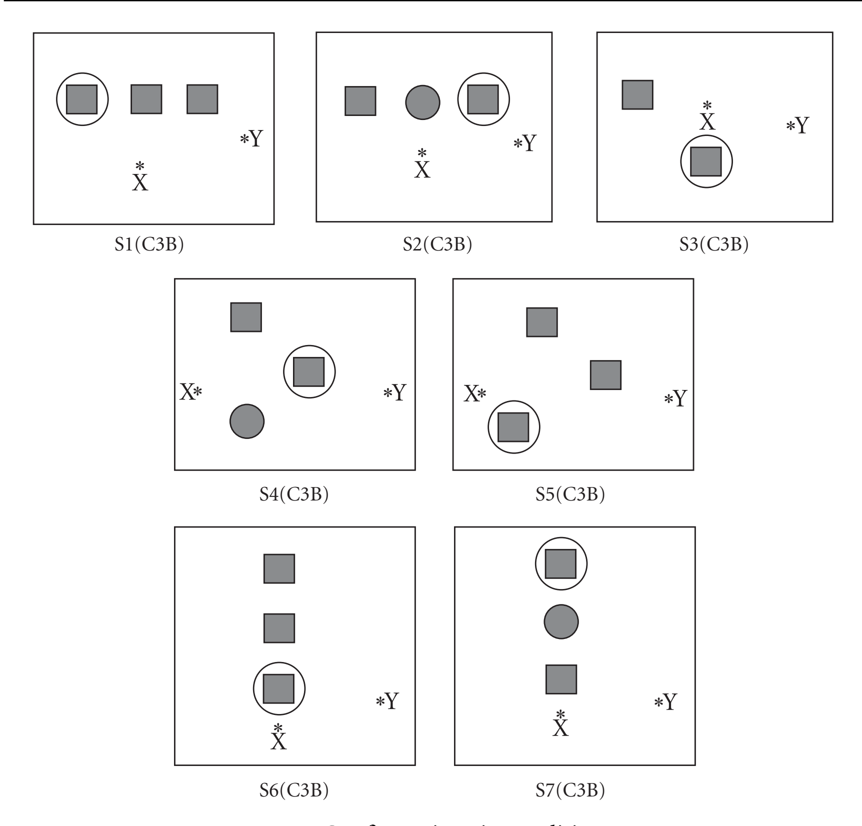


Fig. 8.4. Configurations in condition 3B.

ways, and supporting them statistically. The present work does not include this step, but broadly explores the field, predominantly on a qualitative basis but supported by relative frequencies, in line with established methodology in corpus linguistics. Thus, the data collected in the web-based study were treated as a linguistic corpus rather than conforming to the standards usually established in the field of psycholinguistics. Furthermore, effects of order were not addressed in the present analysis, since they are regarded as a non-trivial additional factor that needs to be treated with specific care. Randomization guaranteed a fairly even distribution of task positions.

8.3 Results of the Empirical Study

In the following, I will analyse the collected data in terms of a range of features on various levels of granularity, starting from overall strategies and ways of referring to the goal object and then moving on to more subtle aspects of linguistic variabil-

ity. Generally, I will refer to the average frequency of utterances across tasks and conditions; for each condition and each task, the mean percentage of each type of expression was calculated. More details including complete frequency lists are available in Tenbrink (2007).

8.3.1 Overall strategy

Speakers of both languages could choose to use strategies other than referring directly to the target object in order to obtain their communicative goal. Specifically in condition 3, a considerable proportion of speakers described the *path towards the goal*, as in 'Walk ahead a few paces. Turn left and walk ahead to the square', rather than simply *identifying the goal square*, as in 'Go to the square on your left', which was also frequently done. Some people did not mention the goal square at all. However, reference to the goal object was the most frequent option in both languages in all conditions.[2]

8.3.2 Ways of referring to the goal object

In some situations, apart from using projective terms, speakers chose other kinds of spatial expressions for referring to the goal object. Among these, the terms that were used most frequently expressed either in-between (such as *middle*) or distance relations (such as *close*). The data show clearly that, if an in-between relation was available for reference, this option was used frequently by speakers. Overall, the average of utterances relying on in-between relations in the relevant scenarios was 28.0% in the German version and 17.2% in English, as opposed to (almost) zero utterances using this relation in other scenarios. Speakers used in-between expressions even in situations where the goal object was not situated directly between two other objects of the same kind, as in S4(C1) (see Figure 8.1 above): they did not necessarily account for digressions from this prototypical configuration linguistically. Similarly, distance-related expressions were used more often in situations exhibiting a clear distance differentiation. However, whether or not they were indeed used for reference was apparently also influenced by other factors, such as the ease of application of projective terms. Apart from in-between and distance relations, further options also used infrequently by both English and German speakers were class names, temporal order, comparative height, compass directions, and clock directions. The percentage of compass expressions used by English speakers was in no situation higher than 3.4% (overall average: 1.4%); only one German utterance contained compass directions at all.

[2] The overall average of goal-based descriptions was 92.5% in German and 78.0% in English; the difference may be due to slight discrepancies in the instructions to the participants (cf. Tenbrink, 2007).

TABLE 8.1. *Syntactic variability and frequencies in English and German*

	German Cond. 1	German Cond. 2	German Cond. 3	English Cond. 1	English Cond. 2	English Cond. 3
N (utt. containing proj. terms)	630	570	652	370	357	231
adjective	61.6	51.6	21.9	51.9	26.1	9.1
noun in prepositional phrase	0	1.2	2.6	31.1	43.4	60.2
handedness term	0	0.5	0	5.9	3.6	0.4
adverb	38.3	42.6	60.3	0.5	10.6	3.5
preposition group	0.2	4.0	15.2	0.8	7.0	25.1
indeterminate form	0	0	0	9.7	9.2	1.7

In the following, I will concentrate on utterances containing projective terms in goal-based utterances. Percentages of frequencies relate to this category only, neglecting the other kinds of strategies just discussed.

8.3.3 Linguistic variability

Whenever projective terms were used for identifying the goal object, speakers employed a considerable range of syntactic variability. However, the variety of linguistic forms was much broader in English than it was in German, especially in conditions 1 and 2 (cf. Table 8.1). In the German data, adjectives and adverbs were used almost exclusively, with some exceptional occurrences of prepositions. In English, usage varied between adjectives, nouns in prepositional phrases, preposition groups, handedness terms, and indeterminate forms (that is, stand-alone terms such as *left* that could either be adjectives or adverbs). In condition 3, the variability of English choices was diminished somewhat. Here, the German and English versions were less differentiated, at least with respect to degree of variety. In some situations speakers settled almost unambiguously for nouns in prepositional phrases in English (and in some situations preposition groups, depending on axis), neglecting other options. German speakers then often used adverbs and in some cases adjectives or prepositions. The main general difference identified by the comparison of the linguistic variability is a much more frequent usage of nouns in prepositional phrases (such as *to my left*/*zu meiner Linken*) in English throughout the data (42.7% for English, 1.3% for German). Also handedness terms (occurring with an overall frequency of 3.8% in English, showing great differences depending on the configuration: the highest frequency was 11.8%) were fairly exceptional in German (0.2% across conditions).

TABLE 8.2. *Overview of syntactic forms in English and German*

Syntactic form	English			German		
	lateral	frontal	vertical	lateral	frontal	vertical
adjective	*right* *left*	front back	upper/ lower bottom/ top	*recht-* *link-*	vorder- hinter-	ober- unter-
noun in prep. phrase: [prep] + [det/ pronoun] +	*right* *left*	front back	bottom top	Rechten Linken		
handedness term	(to the) right-/ lefthand (side)			rechter/ linkerhand		
adverb/ adverbial group; post-nominal modifier	right left	in front/back (straight) ahead/ behind	above below	*rechts* *links*	vorne hinten geradeaus	oben unten
preposition (group)		in front/ back of	above/ below		vor/hinter	über/ unter
indeterminate form	right/left	front/back	above/ below bottom/ top			

Table 8.2 shows a comprehensive list of syntactic forms of English and German projective terms. Here, choices that in the present study occurred particularly frequently are set in italics (only lateral axis, as the other axes were not in focus). With respect to the prepositions used in the construction here called 'nouns in prepositional phrases', it can be added that, in German, only *zu (meiner Rechten)* is possible, while in English, there is a broad and flexible repertory containing at least *to, on, at*, *in*, *towards*, *near*, and *from (the/my right)*, of which the first two were the most frequent in the present scenarios. In English, this kind of construction was preferred in many situations; the only constraint here is that *to* was not used for internal relationships, which occurred most frequently in condition 1 if the picture was used as relatum.

8.3.4 Modifications and combinations of projective terms

In German, lateral adjectives like *recht-* are not available as a superlative; therefore, projective superlatives in German are restricted to the frontal (e.g. *vorderst-*) and vertical (e.g. *oberst-*) axes. In English, in contrast, projective superlatives are straightforwardly used with the lateral axis (as in *leftmost square*). Since most configurations enhanced contrast on the lateral axis, the overall proportion of superlatives was therefore higher in English (11.1%) than it was in German (5.7%). Most German occurrences of superlatives belonged to the 'vertical' scenarios which were not used in the English part of the study: for example 16.7% in S6(C3A) (see Figure 8.3 above). However, not all scenarios allowed for the usage of superlatives, since the application seems to require that the goal element is situated at an extreme position on an axis relative to at least two other elements, which was not the case in all scenarios. This scenario-related fact accounts for the relatively low overall proportion of superlatives in English even though the language itself does not restrict application on the lateral axis the way it does in German.

Furthermore, modifiers of distance were seldom used in German, except in two situations (S2(C3B): 11.1%; S4(C3A): 8.3%) in which specification of distance alone was also often used as a description because the goal element was very close to the interaction partner. In all other situations, the proportion of distance modifiers was never higher than 7.1%. In English, by contrast, distance modifiers (such as *furthest to your left*) were often used to denote an extreme position on an axis, similarly to superlatives, in some scenarios in as many as 35.3% of utterances (S5(C3A)).

With respect to modifiers that render a direction more precise (such as *almost* or *directly [to the right]*), the overall frequency in the German data throughout all scenarios was 15.1%, while in English it was 6.4%. In addition to this quantitative difference, the qualitative distributions of such 'precisifiers' were very different. In the English data, in condition 1 no precisifiers occurred at all, while in German, specifically in S1(C1) and S2(C1), they were fairly frequent (20.9% and 19.2%). In condition 2, in English they predominantly occurred in S5(C2) (22.9% as opposed to frequencies between 2.4% and 4.3% in the other situations), while in German, they were used fairly frequently (between 12.1% and 25.4%) in all situations except S3(C2) (1.8%). Thus, clearly English precisifiers were used in a way different to those used in German. A closer look at the data reveals how. English precisifiers were typically expressions such as *directly* which emphasize a *typical* relationship on a focal axis, and in some cases expressions that point to a *digression* from that axis, such as *a little to the right* and *diagonally to the left*. In German, similar expressions were *direkt (directly)* and *schräg links (diagonally left)*. But in addition, the German data show frequent usage of precisifiers that emphasize an *extreme position* on a focal axis, such as *ganz rechts ('fully right')* or *rechts außen ('externally right')*. This category seems to be missing entirely in the English data. One

candidate for a corresponding expression in English is *all the way on the left*, which occurred once in a path-plus-goal description (in S5(C3B)). Otherwise, such relationships were expressed in English not by precisifying adverbs but by projective superlatives, distance modifiers, and unmodified expressions that may have served the same purpose.

8.3.5 Choice of reference axis

In both languages, speakers preferred the mention of only one axis in most cases (83.3% in German and 84.3% in English across all situations). Two axes were mentioned more often in situations in which one axis was either not discriminative (e.g. S2(C3B): 24.4% in German, 23.1% in English), or if two axes were equal candidates for reference (e.g. S3(C1): 52.0% in German, 48.2% in English). Otherwise, digressions from the prototypical axial direction usually did not induce mention of a second axis. As the overview of linguistic forms for each axis in Table 8.2 above shows, the choice of axis has direct consequences on the range of linguistic variability available. Furthermore, in some cases specific kinds of modifications become impossible, such as German superlatives for lateral terms, as described above. Thus, axes are not neutral with respect to linguistic choice.

8.4 **Discussion**

Generally, participants used a broad spectrum of variability on all scales. The analysis shows that linguistic choices depended heavily on the spatial situation, that is, the presence and placement of other objects and (imagined) persons, and the available kinds of perspective. Therefore, generalized predictions are difficult to formulate on a linguistic surface level. These results are further complicated by a number of language-specific differences, as spelled out in the previous section. Many of the differences identified relate to the structure of the languages involved, as English speakers have a different repertory of forms at their disposal from German speakers. But they also showed different preferences for using and applying this repertory. Notably, for example, nouns in prepositional phrases were a very common way of applying projective terms in English, which was not the case in German, although the form does exist. In German, adjectives were clearly preferred; these are also available in English but they were much less frequently employed. Furthermore, some differences arise with respect to which forms can be used with which kind of axis.

With respect to the usage of modifications of projective terms, the following patterns could be identified. In both languages, projective terms were preferentially not modified if the goal element was the only one on a half plane with

respect to an obvious relatum, even if the spatial relationship between the referent and the relatum was not prototypical (for example, in S3(C3A) the projective term was unmodified in 66.2% of the German and 82.9% of the English utterances; in S4(C2), this was the case for 68.6% of the German and 58.4% of the English utterances). This result represents a strong contrast to previous results on applicability in other kinds of discourse tasks where gradedness of application plays a major role (mostly, 'Where' instead of 'Which' tasks, e.g. Vorwerg, 2001). However, in some situations that may have been perceived as particularly complex by some participants, several kinds of modifications were combined in one utterance (S4(C3B): 29.6% complex descriptions in German and 40.0% in English), and linguistic representations became fairly heterogeneous. This pattern is related to the principle of *redundant verbalization* identified by Herrmann and Deutsch (1976).

Modifications of projective terms occurred in a number of different ways, depending on the spatial situation as well as the language-specific repertory. If two axes were equally good candidates for a spatial description, two projective terms were combined more often than otherwise. If the goal object was at the most extreme position on the lateral axis, in English either superlatives (e.g. *leftmost*) or distance modifiers (e.g. *far right*) were used more frequently than otherwise. In German, precisifying adverbs (e.g. *rechts außen*) were used instead in such situations, which were not used in this way in English. The German usage may partly be due to the fact that, in German, the lateral projective terms cannot be realized as superlatives. In both languages, however, also unmodified projective terms were used to express an extreme position on a spatial axis, similar to the findings of Gorniak and Roy (2004). If the goal object was particularly close to the relatum, in both languages a distance or precisifying term could be used in addition to the projective term. If the goal object was situated roughly between two other objects, an in-between relation could be expressed in addition to the projective term (e.g. *middle right*). If the goal object was situated directly on a focal axis, a precisifying modifier was frequently used to emphasize this relationship in both languages (e.g. *the one right in front of you*).

Thus, the present research has led to the identification of different preferences for syntactic forms for speakers of the two languages in relation to subtle differences in the spatial configuration. Since former investigations have typically focused either on the choice of reference systems or on the variability in language structure, the results gained here represent a new approach towards identifying and contrasting application preferences in a referential identification situation in two languages.

With respect to the underlying principles and preferences for applying projective terms, some tendencies could be identified that hold for both languages alike, abstracting away from language-specific differences. These are in accordance with previous research, and the general strategies worked out in detail in Tenbrink

(2005) for English were confirmed for German as well. Three major principles, at least, seem to be at work, influencing speakers' choices. The principle of *contrastivity* ensures that the goal object can be identified among the competing objects. The principle of *minimal effort* leads, on the one hand, to the omission of information that is redundant or easily inferable and, on the other hand, to linguistic and conceptual choices (including projective terms versus other kinds of spatial expressions, relata, and spatial axes) that enable referential identification with a minimum of additional information encoded as linguistic modifications. The principle of *partner adaptation* seems to be chiefly responsible for the choice of perspective.[3] All of these principles that together explain the speakers' preferred choices to a high degree are in accord with earlier, more generalized findings in the literature (e.g. Herrmann and Deutsch, 1976; Clark and Wilkes-Gibbs, 1986; Schober and Brennan, 2003). Schober's results (this volume) add to these principles by further exploring how the principle of partner adaptation is active in dialogue between partners with mismatched abilities.

These results support the assumption (discussed in more detail in Tenbrink, 2005) that web-based studies, in spite of their obvious drawbacks (see, for example, Reips, 2002), can be useful for the analysis of speakers' choices and preferences in a discourse task that is not exactly natural but still more realistic with respect to the freedom to select expressions than a controlled laboratory situation. Therefore, the conclusions drawn here that pertain to the comparison of English and German may serve as useful starting positions for focused and more controlled empirical investigations directly testing the hypotheses formulated. Without such prior exploration, there is a certain danger of exploring expressions in the laboratory that are not naturally used by speakers in corresponding real-life situations, raising questions about their ecological validity. Specifically, confederate studies which address alignment effects may profit from focusing on those terms that are naturally preferred by speakers, as identified by free production studies such as the present one. Also, the results are informative for the development of dialogue systems capable of handling potentially ambiguous reference resolution tasks (e.g. Dale and Reiter, 1995; Wyatt, 2005).

8.5 **Conclusion**

The present study has offered an exploration of natural language produced in an open (though artificial) setting by unbiased native speakers of English and German. This has led, on the one hand, to an assessment and comparison of the diversity in the linguistic choices of speakers of the two

[3] In German as well as in English, typically the interlocutor's point of view was adopted if available. In German, such choices were made explicit somewhat more often (see Tenbrink, 2007, for details).

languages in a range of situations allowing for different interpretations and viewpoints, and on the other hand to the identification of systematic conceptual and communicative principles and strategies underlying speakers' choices, which overwhelmingly seem to be independent of whether English or German was used. The results partly echo earlier general results specifically for a less well researched type of discourse task, namely contrastive spatial reference, and partly point to interesting language-specific differences in speakers' linguistic preferences that have not been addressed systematically before.

9

Explanations in Gesture, Diagram, and Word

BARBARA TVERSKY, JULIE HEISER, PAUL LEE, and MARIE-PAULE DANIEL

9.1 Beginning: Characterizing Explanations

People are constantly explaining things to one another. Parents explain to children how to build a tower of blocks or why they need to go to bed. Friends explain to each other how to find their homes or why they were late. Teachers explain to students how nerve conduction happens or why World War I began. In person, explanations, in common with most face-to-face communication, are typically multimodal. Not simply talk, explanations include gestures, props, and often diagrams (see for example Engle, 1998). Each mode has relative advantages and disadvantages, and they work in concert, complementing and supplementing one another (Clark, 1996; Goldin-Meadow, 2003). Frequently, explanations are restricted to a single mode. Giving directions over the phone limits them to words. Providing instructions to operate a camera or assemble a piece of furniture to international consumers limits them to diagrams. Although gesture is often used alone for brief interchanges, like signalling to a partner a desire to leave a party, it is less likely to be solely used in complex explanations. Those who have got lost in a country whose language they don't know learn the power of gesture alone. Limiting communicators to a single mode requires completeness of that mode. This reveals the structure of explanations and the parallel ways each mode expresses that structure.

Not all explanatory tasks readily lend themselves to words, diagrams, and gestures. Fortunately, two of the most common kinds of explanatory tasks do: navigation and construction. One of each was selected in order to investigate the structure and semantics of explanations. The navigation task was to communicate how to get from one place to another. The construction task was to communicate how to put together a piece of furniture, a TV cart. The data analysed here were gathered from several experiments differing somewhat in method. In all cases, participants first learned or already knew the specifics of the task and then provided instructions so that another person could perform the task. In some experiments, participants were asked to provide verbal instructions,

in others, diagrammatic instructions, and in others, explanations using gestures with props, either with or without words. In still other experiments, participants could use various combinations of modes, such as diagrams and words or gestures and words.

The two tasks, constructing an object and learning a route, are representative of the kinds of tasks people learn in ordinary as well as technical situations. Both tasks require explanation of actions in space. They both stipulate the arrangement of parts in a particular spatial-temporal configuration. As such, they are similar to tasks such as making pot-au-feu or operating a voting machine or understanding how the heart works or performing surgery or maintaining a power plant. At the same time that they are representative of tasks requiring knowledge of the spatial-temporal organization of parts, they are readily learned by novices in a laboratory session. Route learning and assembly tasks have been studied extensively from different perspectives (Allen, 2000; Denis, Pazzaglia, Cornoldi, and Bertolo, 1999; Novick and Morse, 2000; Shi and Tenbrink, this volume; Striegnitz, Tepper, Lovett, and Cassell, this volume; Tversky, Agrawala, Heiser, Lee, Hanrahan, Phan, Stolte, and Daniel, 2007). Because these tasks entail explanations of the spatial configuration of parts in a temporal sequence, they are likely to elicit numerous gestures (e.g. McNeill, 1992; Krauss, Dushay, Chen, and Rauscher, 1995; Wagner, Nusbaum, and Goldin-Meadow, 2004).

Like narratives, explanations have a discourse structure consisting of a beginning, middle, and end. For explanations, the beginning is an introduction, the middle, a step-by-step set of procedures (some with qualifications and embellishments), and the end, an indication that the task has been completed. This discourse structure has been observed in spontaneous verbal instructions (e.g. Denis, 1997) and in diagrams as well (Tversky *et al.*, 2007). The semantics crucial for explanations includes referring expressions for objects (or object parts) and for actions. Objects are typically static whereas actions are dynamic. Of interest here is how the discourse structure and the semantics are expressed visually, in diagrams or gesture.

Of the three modes of communication, words are perhaps the most common. Although words are purely symbolic, there are a great many of them, allowing nuanced expression of concepts, actions, and relations. Words can do a number of things that are difficult to do with pictorial or visual media: for example, words can be used to qualify, to negate, and to hypothesize. Diagrams are also familiar modes of communication, most certainly maps for navigation and diagrams for construction. Diagrams have a number of advantages over words. In contrast to words, diagrams bear visual similarity to what they are meant to communicate, both objects and actions. Diagrams can depict objects and object parts, their structural relations, and even their manner of assembly to convey construction. Diagrams can indicate landmarks by names or icons and can schematize the turns that constitute a route by a turning line to convey a route.

Gestures, like diagrams, are a visual mode of communication. They serve a number of roles in discourse as well as in thinking, that is, for those using gestures as well as for those observing them (Goldin-Meadow, 2003; Kessell and Tversky, 2005; Krauss, 1998; McNeil, 1992; Sowa and Wachsmuth, this volume). Several taxonomies of gestures have been developed (e.g. Ekman and Friesen, 1969; Kendon, 1988; McNeil, 1992; Rimé and Schiaratura, 1991). Some gestures serve to structure the discourse, for example using the hands to separate arguments, often accompanied by 'on the one hand' and 'on the other hand'. Others, termed emblems, have conventional meanings like words, for example waving goodbye or the sign for *OK*. Here the focus is on two other types of gestures, those central to explanations; specifically, the deictic and iconic gestures that carry semantic meaning and gestures that form the discourse structure characteristic of explanations. On the whole, *deictic* gestures point to or indicate things in the environment; *iconic* gestures resemble what they are meant to convey. Some gestures function by interacting with things in the world, by referring to and acting on things in the environment. In the present cases, the things are maps for route directions and object parts for assembly instructions. Gestures, then, are inherently both embodied and situated.

Partly because they are both situated and embodied, gestures can support or even convey explanations in a rich set of ways. Some gestures, like *deictics*, can refer to particular aspects of a situation rather than others, directing and focusing attention on the critical aspects of a situation. Unlike most words, gestures, specifically *iconic* gestures, may bear physical similarities to the things they refer to: for example, illustrating a T intersection by making a T with the hands. This physical resemblance may make gestures easier to interpret and easier to remember. Note that the T may be conveyed in gesture in different ways: by the whole hands, by two fingers, by writing a T in the air. At the level of semantics, the particular hand and finger positions can vary considerably and still convey the same meaning. Because gestures occur in a spatial medium, and because people think and talk about many abstract concepts in spatial metaphors, gestures can also bear *metaphoric* relations to the things they represent. For example, when telling a friend that she and another friend had had little recent contact, one person said, 'We've grown apart', while separating her two hands. The metaphor of spatial distance representing psychological distance was expressed symbolically by the words, but concretely by the hands. The actual distance between the hands can be used to convey the degree of separation, an aspect of communication not conveyed in the words. Gestures share both these qualities, iconicity and metaphoricity, with diagrams. Part of the effectiveness of diagrams derives from physical and metaphoric similarities to the things they represent (see for example Tversky, 1996, 2001). When gestures are redundant with speech, they provide a second way of encoding information, in addition to words. For memory, two codes are better than one (e.g. Paivio, 1986). Moreover, for concepts that can be depicted, a pictorial code is superior to a verbal one, presumably because of the resemblance of pictures to the things

they represent. By analogy, iconic gestures should have an advantage over arbitrary words.

Gestures can do more than depict, they can also enact. When describing how he swerved his car to avoid a potential collision, one speaker swayed his body deeply to show the swerve. Notice that it was the car that swerved, not the body, so the swerving was metaphoric, though both passengers and drivers are known to swerve when concerned about collisions. Swerving and other demonstrations of action, such as showing another's strutting or demonstrating a knot, *embody* the knowledge they are meant to convey, coding it motorically as well as pictorially. Motor codes, like pictorial codes, are also known to augment memory (e.g. Engelkamp, 1998). Finally, gestures can situate knowledge in the relevant context, by using actual or virtual props. When describing how to make a pendulum from a rope hanging from the ceiling, one speaker first used his hand to indicate the position of the rope, then to indicate the weight to be tied on the end of the rope, then to show the tying, then to indicate grasping the rope and setting it in motion, and, finally, to indicate the swinging of the rope (Kessell and Tversky, 2005). It is remarkable how many different meanings the hand was used to convey, smoothly and effortlessly. In situating the knowledge, gestures can create a mental model of the objects and the space as well as of the actions (see for example Emmorey, Tversky, and Taylor, 2000). Because they are spatial and (in the case of gestures) temporal, diagrams and gestures can convey models, structural, behavioural, or causal, more directly than words.

9.2 **Explaining How to Put Something Together**

9.2.1 Production of assembly explanations

In the experiments eliciting explanations of assembly, participants first assembled a TV cart on their own, using the photograph on the box as a guide (see Heiser, Phan, Agrawala, Tversky, and Hanrahan, 2004). The TV cart was 17″ x 25″ × 21″ in size, and consisted of two sideboards, an upper shelf, a lower shelf, a support board, pegs for attaching the support board, screws, screwdriver, and wheels. After assembling the TV cart, participants in the various studies were asked to design instructions explaining how to assemble it. Some groups produced diagrams and words, some only diagrams, some only words; still other groups made explanatory videos as they reassembled, free to use both speech and gesture, while other participants were asked to make explanatory videos as they reassembled, but told that the videos had no sound track, and might be used by people who did not speak English. Here, the nature of the verbal, diagrammatic, and gestural explanations is analysed. The focus is on gestures used in explanations, as that aspect of the work is new.

9.2.2 Gestures

Narrative structure. Explanations, like other forms of discourse, have a narrative structure, notably, a beginning, middle, and end. Think of recipes: they begin with the name of the dish to be cooked, and then list the ingredients, typically in order of use. They continue with the list of procedures in the order to be followed. The end of a recipe is the outcome, typically the number of people it will serve. In our data, explanations of TV cart assembly and route directions also frequently had a discourse structure. That structure was often carried by gestures, especially in the 'gesture alone' group. A variety of gestures served narrative roles, some used by some participants, others by others. Most participants began with some sort of an introduction. Many waved to the camera by way of introduction. Others began as a recipe begins, by presenting the parts to be used to construct the whole object. The narratives continued with a step-by-step explanation. Frequently, the steps themselves had an internal structure, a beginning, middle, and end. Finally, the explanations often used gestures to signal the end of the narrative: that the task had been accomplished. Some of the devices used to carry the narrative and convey the explanation are described in the following sections, beginning with presenting parts. Parts were often presented as part of the opening of the entire explanation and also as the opening of each step.

Introduction: presenting parts. Presenting parts typically took one of two forms, exhibiting and pointing. Both of these involve *deictic* gestures—whose primary function is to point to or indicate something in the environment. Typically, the large parts were exhibited by holding them up to viewers, and the small parts were pointed to. Parts were presented not only as an introduction to the entire task but also as an introduction to each step.

Conveying steps. After presenting the parts, the explanatory narrative typically continued by providing each assembly step in turn. Discourse structures are often hierarchical, and these were no exception. At the more general level, the discourse structure of procedural explanations has a beginning, a step-by-step specification, and an ending. But each of these discourse parts may have subparts, especially the steps that constitute the middle. Each step had a beginning, middle, and end. Those who could speak used language to convey narrative structure. For those discouraged from speaking, a variety of gestures served to convey narrative structure. Gestures were used to mark the beginning and end of each step, and to explain each step by demonstrating it. The beginnings of steps were often marked by holding up a finger to indicate the step number, and the endings by a gesture such as 'OK' or 'thumbs up', or flattening the two hands palms down and moving them away from the body and outwards as if to push away to indicate 'done'. Perhaps because of the relative ease of marking steps in language, using words like 'next', 'after that', or 'now', speakers marked steps explicitly more often than those restricted to gesture. Gestures were also used at the beginnings of steps to present the parts that would be used for that step.

Models of structure and action. Another way that steps were introduced was by previewing the structural change or the action needed to accomplish the structural change. In gesture, this was accomplished by a gestural model. A little-noted feature of gestures is that sequences of related gestures can be used to convey a mental model (Enfield, 2004; Engle, 1998; Emmorey, Tversky, and Taylor, 2000). For example, in describing environments, participants used as many as 15 related gestures in a sequence to locate landmarks in space, maintaining a spatial display 'drawn' in the air, with consistent size and position (Emmorey *et al.*, 2000). In order to explain the assembly of the TV cart, participants often made models using gestures, frequently prior to an assembly step. For TV cart assembly, the models often used parts of the TV cart as props. The models were of two types: structure or action. For structure models, explainers used their hands, sometimes with the parts, to show the desired structural relations among the parts, for example using the palm flattened horizontally to indicate the top of the cart, and both hands flattened vertically to indicate the sideboards. For action models, explainers used their hands, sometimes with the parts, to show the actions required to achieve the desired structure; for example, holding real or imaginary parts and moving them in place. Both structure and action models used a combination of *deictic* gestures that point to things in the environment and *iconic* gestures that bear resemblance to what they are communicating. The demonstrations of action, for example, were iconic gestures, as were sequences of gestures showing the shape of the object.

Ending. Participants often took care to indicate when the task was done, even though the completed cart was a self-evident ending. One way they did this was by a gesture of presentation. In contrast to the step-ending gesture, the palms of the two hands faced upwards as the hands unfolded before the completed cart. Other endings included an 'OK' sign or thumbs up. The sequence of gestures, then, can tell a story, a particular kind of story, namely an explanation. In this case, gestures were used to indicate a beginning, often including a greeting, commonly with a presentation of the parts to be assembled. Gestures marked assembly steps, and introduced them with a demonstration of the action to be performed. They marked ends of steps. Gestures can also end the story, often by presenting the completed object.

It should be remembered that not all of the participants used gestures to indicate all the components of the narrative. Gesture narratives were more frequent among explainers restricted to gestures. For them, the gestures carried the entire message. Those using only gestures acted as if they needed to create a coherent and integrated set of gestures in order to convey the assembly task. So, for example, if they used a gesture to introduce the task, they also tended to use a gesture to end the task, often gestures that were related, such as waving 'hello' at the beginning and waving 'goodbye' at the end. Just as in speech where words are related and integrated, so were gestures in those not allowed to speak. For those who could

speak, the language could carry most if not all the message, so the gestures were more optional, and did not have to be integrated into a coherent set.

Action gestures are also actions. As such, they *embody* the information that is crucial for task performance. They also *demonstrate* the information that is crucial to performance. In addition, the action gestures *situate* the information necessary for task performance: they were formed with respect to virtual or actual parts and performed with respect to a virtual or actual object. These properties are likely to render gestures as especially important to communication and comprehension, especially for information literally or metaphorically about spatial relations and about action. Gestures can directly map spatial relations to be conveyed, such as the structure of an object; they can also directly map actions to be comprehended, such as moving or rotating parts. Diagrams also have a special status in communication and comprehension because they too can map spatial relations in the information to be conveyed to spatial relations on paper. Diagrams also embody techniques for conveying actions, notably in the form of arrows (see for example Tversky, Zacks, Lee, and Heiser, 2000; Tversky, Heiser, Lozano, MacKenzie, and Morrison, 2007). However, the techniques for conveying action using gesture are more direct, just as the techniques for conveying structure are more direct in diagrams. Now we turn to examine briefly how explanations are accomplished in the visual medium of diagrams.

9.2.3 Diagrams

The diagrams explaining construction produced by users also had a narrative structure, especially those drawn by people high in spatial ability (Tversky *et al.*, 2007). Like a recipe, the beginnings of explanatory diagrams were often a list of ingredients, namely the parts to be used in construction of the TV cart. As for the explanations in gesture, diagrammatic explanations typically had a middle consisting of the sequence of steps needed to assemble the object. The better diagrams showed each step in the perspective needed for assembly, and embellished the depictions of the object parts with arrows and guidelines indicating how the parts should be moved into position, that is, the better diagrams showed the actions needed to put the parts together, not just the structure of the parts. Each new step was a new part to be attached. Finally, diagrammatic explanations often used a deliberate ending. In many cases, participants drew lines surrounding a sketch of the completed TV cart, like rays around a sun.

9.2.4 Words

Just as for explanations relying on gestures and explanations relying on diagrams, explanations relying on or using words had a narrative structure (Daniel, Tversky, and Heiser, 2006). Most participants introduced the task, many as with

the gestures and diagrams, by a listing of the parts. Others began by providing a structural mental model of how the parts fitted together to make a whole. Still others gave general advice. Some participants used no special beginning, but rather started right in with the step-by-step instructions. The middle, as for gestures and diagrams, was a hierarchical set of step-by-step instructions that specified the actions and subactions to be taken on each object part, that is, actions on objects or object parts. The instructions also often included qualifications; for example, perceptual details that helped to identify the relevant part, or action details that specified the manner of action. The steps were often explicitly marked, for example, by 'first', 'next', and 'finally'. Endings varied, sometimes simply saying the task was now finished, sometimes adding a suggestion about how it could now be used or suggesting that the user could be proud of finishing.

9.2.5 Assembly instructions and mental model of assembly

The three modes of communicating share a number of features that suggest that they also share the same underlying discourse structure and the same underlying mental model of assembly. Spontaneously produced explanations, like stories, typically had a narrative structure: they were not merely a list of a steps but an integrated list sandwiched between a beginning and an end, an introduction to the task and an indication that the task was completed. The heart of the instructions was the set of actions on objects, whether expressed gesturally, diagrammatically, or verbally. The actions on objects were the goals and subgoals that constituted the task rather than motions. That is, rather than 'slide the shelf horizontally', participants said or showed the goal of the sliding. Typically, each step involved a new object part. Thus, the mental model of assembly is a hierarchical set of actions and subactions on objects or object parts. These insights have come from people's production of procedures for carrying out an organized set of actions with a beginning, a middle, and an end, constituting an accomplishment or achievement. The processes that complement production are perception and comprehension of organized sets of actions, having beginnings, middles, and accomplishments, termed *events*. Research on event perception and cognition have revealed the same mental model as the present research on production of assembly instructions: that is, events are perceived and conceived to be a hierarchical set of actions on objects accomplishing goals and subgoals (see for example Tversky, Zacks, and Martin, 2008; Zacks, Tversky, and Iyer, 2001).

9.3 **Explaining a Route**

Wayfinding is a skill that precedes humankind, indeed, remarkably so. Think of migrating butterflies, ants, fish. And *Lassie Come Home*. What humans have added is ways to communicate routes to others, including describing them, sketching

them, or gesturing them. The structure of routes, whether described, gestured, or depicted, consists of a sequence of actions, typically turns, on paths at landmarks (Denis, 1997; Tversky and Lee, 1998, 1999). How are these and other components of routes communicated in each mode?

9.3.1 Gestures

Participants were given four maps (London, Paris, Palo Alto, San Francisco Bay Area), each with a start point and an endpoint of a highlighted route. They were asked to devise a route from the start point to the endpoint, and to explain that route to a camera while making the map visible to the camera so that someone else viewing the video could find their way. Some of the participants were free to use gesture and speech and others were told that the sound track was off because viewers might not understand English.

In explaining how to assemble the TV cart, communicators used deictic gestures to indicate parts and iconic gestures to demonstrate actions. For explaining routes, communicators used both deictic and iconic gestures. Deictic gestures were frequently used to point to places; iconic gestures were often used to demonstrate action, notably turns. Here, communicators demonstrated action using their hands rather than their feet. Since communicators had maps, they could gesture on the map by pointing to landmarks where turns occur and tracing the paths. Gestures like tracing the path are iconic with respect to the map but metaphoric with respect to the action.

Narrative structure. As for assembly explanations, the gestures used in route explanations, especially by those restricted to gesture, often had a narrative structure. Those restricted to gesture tended to produce an integrated, related set of gestures. For a *beginning*, many communicators greeted the recipients by waving; others pointed vigorously at the start point. Some pointed first to the viewer, then to the start point, a way of orienting the viewer in the map, analogous to 'you are there' indicators. The *middle* was typically a step-by-step explanation of the route, consisting primarily of deictics to indicate locations, landmarks, or turning points and iconics to indicate form of paths or turns.

Metacomments. Explanations often include a variety of metacomments in addition to the step-by-step information. Remember that such comments were common in the verbal instructions for assembling the TV cart. Gestures were used in these. For example, on some occasions, gesturers explicitly chunked the information, first showing a few turns one by one, and then reviewing a group of them. Gesturers also made off-map gestures to clarify components of the route. They sometimes included iconic gestures specifying landmarks, such as a series of gestures that conveyed a hotel, first by outlining a building, then by miming its function, for sleeping. They also used off-map gestures to clarify a route, switching from the survey or overview perspective of the map

to the route or embedded perspective of the traveller (cf. Taylor and Tversky, 1992).

Endings. Finally, gesturers ended their narratives most typically with an 'OK' or thumbs up or presentation gesture, both palms out and upwards as if presenting the explanation to the viewer.

9.3.2 Diagrams

A number of years ago, university students were approached outside a dormitory just before dinner hour and asked if they knew the way to a local restaurant (Tversky and Lee, 1998, 1999). If they did, they were asked to either sketch a map or to write directions to the restaurant. Both sketches and directions were analysed. A single diagram was used, and the beginning, the start point, the middle, or the set of turns at landmarks, and the ending, the restaurant, were marked. Notably, although the sketch maps could have been analogue, reflecting actual distances and angles, they were instead schematic; that is, they did not show exact metric relations, either distance or angle. They left out streets and landmarks not directly involved in the route, retaining primarily the streets and landmarks of the route. Turns, irrespective of actual angle, were shown as right angles. Short distances with complex actions were relatively enlarged whereas longer distances with no change of action were relatively reduced. Thus, the diagrams were distorted in ways that made the route more readily apparent.

9.3.3 Words

The language of route maps parallelled the sketch maps. Denis and his colleagues (Denis, 1997; Denis, Pazzaglia, Cornoldi, and Bertolo, 1997) have analysed the language of routes, and our results replicate theirs. Those directions were given in response to a question about how to get to a destination, so the start points and endpoints, the beginnings and ends, were already established and did not need to be communicated. Nevertheless, route directions often had an explicit *beginning*, orienting the traveller. The *middle* consisted of iterations of locating a landmark and specifying an action. Landmarks were typically referred to by names, names of buildings or of street intersections. Actions, typically turns, were conveyed by terms such as 'turn right', 'take a right', or 'make a right'. Other actions, typically progressions, were referred to as 'go down' for straight paths or 'follow around' for curved paths. Note that the descriptions of actions, parallelling the sketch maps, did not specify the exact angle of turn or the exact distance of procession. As for the verbal assembly instructions and the gestural route directions, the verbal route directions often included redundancies and metainformation, such as using more than one perspective, route, and survey, local summaries of set of turns, specifications of landmarks and warnings of pitfalls. Finally, route directions

often explicitly marked an *ending*, such as 'There you are' or 'Now you've arrived'.

9.3.4 Route directions and mental models of routes

As for assembly, the data from the three modes of communication of routes suggest a common underlying mental model for routes: a sequence of turns or actions at landmarks where distance and angle are schematized, that is, not specified. Correspondingly, communications of routes, whether by gesture, diagram, or word have beginnings that orient a recipient, middles that provide a step-by-step set of actions at landmarks, and endings that indicate arrival.

9.4 **End: Modes of Explanation**

Explanations are a common form of discourse. Like stories and expository prose, explanations, the present kind of discourse, are structured, with beginnings, middles, and ends. Yet in other ways stories differ from explanations. Whereas stories typically have a narrative voice, explanations typically do not. Stories are usually about life events of people, explanations about possible events of systems, objects, or people. As such, stories are typically imbued with emotion whereas explanations are not. Stories tend to be episodic, explanations semantic, in Tulving's (1972) sense of the terms. Here we have examined two paradigmatic kinds of explanation, how to put something together and how to get from here to there.

We have compared and contrasted spontaneous explanations that use one or more communication mode, gesture, diagram, or word. The explanations were of actions in space, putting something together or navigating a route. Each involved objects, landmarks or parts, and actions performed on or at them. The exact form of the beginnings, middles, and ends varied, depending on the affordances of the communication mode adopted. For each kind of explanation and each mode, the beginning presumed an initial state of ignorance and a goal: to learn the task. Correspondingly, beginnings for the assembly task sometimes greeted the recipient of the communication, sometimes overviewed the task, and sometimes presented a menu of parts. For the route task, beginnings often greeted the recipient or oriented the recipient in the environment. The middle consisted of a set of steps or procedures, often explicitly marked, that were actions on objects for assembly or actions at landmarks for routes, that is, actions with respect to a spatial configuration. Finally, explanations ended with some indication that the task was completed.

All modes—gesture, word, and diagram—appear to serve thought as well as communication. People have been known to talk to themselves, to think aloud, especially in the shower. Designers, mathematicians, scientists are often at a loss

without the proverbial cocktail napkin to sketch on. Indeed, it has been proposed that designers use sketches to hold conversations with themselves (Schon, 1983; also Goel, 1995; Goldschmidt, 1991). In fact, designers, especially experienced ones, get new ideas from inspecting their own sketches; they 'see' new and unintended relations, patterns, functions that emerge from the sketches (Suwa, Tversky, Gero, and Purcell, 2001). Interestingly, when architects are asked to design while blindfolded, they gesture profusely, as if the gesturing replaces the sketching they normally do (Z. Bilda, personal communication, 2005). Diagrams are thought to be effective for communicating a broad range of ideas for a number of reasons (Tversky, 2001). Diagrams use elements that may bear a physical or metaphorical relation to what they represent, for example, icons in software or airports. They use spatial relations to convey relations that are directly spatial, such as maps, or metaphorically spatial, such as degree of productivity or attractiveness. Communications that are iconic or metaphoric are more direct than purely symbolic communications. Indeed, diagrams (and gestures) can often be understood in situations where the local language is not. Diagrams encourage completeness of thought: unlike words or gestures, they do not readily tolerate ellipsis (see for example Tversky and Lee, 1999).

Gestures, like diagrams, are effective in part because their relationship to meaning is more direct, less mediated. In addition, and in contrast to words and diagrams, gestures can embody the knowledge they are meant to convey. This is particularly true for action information, as gestures are frequently mini-actions, such as pushing and pulling or placing. There is evidence that embodiment itself is privileged for learning action information, even when retrieval is verbal (Engelkamp, 1998). Finally, gestures situate knowledge in the world in which it will be used. They do this by pointing to the objects and the places that they hold or will hold with respect to other objects. In the present cases, situating knowledge was facilitated by the props provided, the object parts or the paper map. However, in cases without props, gestures are often used to establish virtual or proxy props, and other gestures act on them. For example, in describing environments without props, people frequently created a map in the air, adding landmark after landmark to the appropriate place in the imagined map (Emmorey *et al.*, 2000). Situating knowledge embeds it in a rich structure, part of which may already be known, that provides a scaffold for new information. As for many cognitive phenomena, then, the advantage of gesture for both communicators and recipients has more than one source. We have proposed four—that (deictic) gestures draw attention selectively to the critical aspects of the message; that (iconic and metaphoric) gestures bear a literal or figurative likeness to what they convey; that gestures embody knowledge; and that they situate knowledge in the world; but there may be others.

Gestures have advantages, but they are not, per se, a complete language, though they can evolve into one (see Goldin-Meadow, 2003). Words, of course, are, though it must be noted that spontaneous spoken words typically lack the

completeness that written text or prepared talks have. For some tasks, especially explanations of routes and assembly, diagrams are complete and sufficient. Yet, despite completeness, combining these modes seems to create the most effective explanations, as each has something special to contribute. People have a great many stories to tell, expositions to relate, and explanations to convey; fortunately, they have a rich set of means of expression to do so.

Acknowledgements

Correspondence concerning this article should be addressed to Barbara Tversky, Columbia Teachers College, 525 W. 120th St., New York, N. Y. 10027, U. S. A. E-mail: btversky@stanford.edu. Thanks are due to Herb Clark for his insights. The present analysis draws on work with Pat Hanrahan, Maneesh Agrawala, Doantam Pahn, Sandra Lozano, and Chris Stolte. Portions of the research were supported by Office of Naval Research Grants N00014-PP-1-0649, N00014011071, and N000140210534, NSF Grant REC-0440103, and an Edinburgh-Stanford Link grant to Stanford University.

10

A Computational Model for the Representation and Processing of Shape in Coverbal Iconic Gestures[1]

TIMO SOWA and IPKE WACHSMUTH

10.1 Introduction

When describing object shape, humans usually perform iconic gestures that coincide with speech (that is, they are *coverbal*). Iconic gestures are semantically related to the content of speech (co-expressive), but express meaning in a different way (McNeill, 1992, 2005). They unfold in time and space and present images, marked by a similarity between the gestural sign and the described object. In contrast, speech unfolds only in time and presents a stream of symbolic signs arbitrarily connected to meaning. Still, both modalities, the verbal and the gestural, are assumed to form an integrated system of communication (Bavelas and Chovil, 2000; Clark, 1996) with a common origin or 'idea unit' underlying the production of co-expressive speech and gesture fragments (McNeill, 1992, 2005; Kendon, 2004). Due to their inherently space-bound nature, iconic gestures may easily depict content that is difficult to describe using words alone. This iconic content is picked up and processed by listeners as recent studies on gestural mimicry and neuropsychological findings suggest (Kimbara, 2006; Kelly, Kravitz, and Hopkins, 2004). Though the expressive potential of iconic gestures in human–human communication is generally acknowledged, few systems make use of it in human–computer interaction. Instead, the development of comprehension systems that take the non-verbal component of natural communication into account has focused much more on pointing and symbolic gestures.

Our main goal is to build computational models for the representation and processing of spatial language including shape-related gestures. We employ empirical studies on communicative behaviour as the main information source for our modelling efforts. In that respect our contribution is in line with the empirically-based work of Shi and Tenbrink (this volume) on dialogue design

[1] This chapter is an elaborated version of a paper first published in K. Opwis and I.-K. Penner (eds.) (2005), Proceedings of KogWis05, Basel: Schwabe Verlag.

for an instructable wheelchair and Striegnitz, Tepper, Lovett, and Cassell (this volume) who focus on spatial knowledge representation for the generation of verbal and gestural route descriptions by an artificial agent. In this chapter, we concentrate on the way people use iconic gestures in descriptions of object shape, and we present an approach to employ such gestures in comprehension systems for spatial language. We show how to build up semantic representations for multimodal referential expressions like the noun phrase *a longish bar* + *<iconic gesture>* in which the adjective, the noun, and the gesture together form a composite signal (in the sense of Clark, 1996) specifying an object. Application areas include natural language interfaces for autonomous systems (e.g. mobile robots), virtual construction applications, and virtual design.

The chapter is organized into three parts. First, in section 10.2, we describe an empirical study on the use of iconic gestures and speech in shape descriptions. Second, in section 10.3, we describe a formal representation for multimodal shape descriptions that captures the content of shape-related iconic gestures as well as shape-related adjectives and nouns. Finally, in section 10.4, we sketch the application of the representation in a multimodal system for gesture and speech comprehension.

10.2 Gesture and Speech in Shape Descriptions

In order to examine the morphology and the semantic aspects of shape-related coverbal gestures, an observational study was conducted which is described in more detail in Sowa and Wachsmuth (2003). A total of 37 participants was asked to describe five different stimulus objects (Figure 10.1). The objects were projected with a video data projector on a wall-size screen. It was decided to use magnified projections of the original parts since large stimuli were assumed to evoke larger and clearer gestures. The height of the cube in Figure 10.1, for instance, was about 80 centimetres on the projection screen. The original parts were not shown to the subjects.

The participants were told that their descriptions would be videotaped and shown to another person afterwards. They were instructed to give their description in such a way that the person watching the video recording would get an idea of the appearance of the objects. It was mentioned that the hands could be used for the descriptions, but hand gestures were not enforced nor were the



Fig. 10.1. Stimulus objects used in the study.

descriptions in any other way restricted. The stimulus objects were computer graphically-generated parts of a toy construction kit. Two pairs of objects, the stylized screws and the bars, were quite similar. Though their basic shape was nearly identical, they differed in their sizes and proportions. That way, effects of size and proportions on iconic gestures could be examined in isolation. All gestures judged to express shape-related content were transcribed with respect to spatiotemporal features, that is, their form, and the corresponding elements of meaning. The annotated corpus comprises 383 gestures. The analysis of verbal information in the corpus relies on the concept of *lexical affiliates* which could be single words, multiple words, or phrases to which gestures semantically relate. For each gesture transcribed, its lexical affiliate was determined independently by three coders. Only those words or parts of speech rated as lexical affiliates by at least two coders were included in the analysis.

10.2.1 Gesture types

In order to systematize the corpus, gestures with a similar relation between form and meaning were grouped together yielding 84 different gesture kinds. The form–meaning relation was considered similar if identical spatiotemporal features had been used to express the same semantic properties. For instance, one gesture kind is marked by the distance between the tip of the thumb and another fingertip expressing object extent. Several variants of this form–meaning relation were observed. The finger opposing the thumb's tip could be the index, the middle, or even the ring finger, and the extent could be displayed horizontally or vertically. However, all of these concrete instances share the general feature that two points in space are defined by opposing fingertips, expressing the property of extent. Each gesture kind can be represented by a prototype which is an idealized realization of the form–meaning relation (see Table 10.1 for the most frequent gesture kinds). Four general gesture types can be distinguished as given below.

10.2.1.1 *Dimensional gestures*

The largest group is characterized by representing an object's outer dimensions via delimiting or enclosing. Such gestures may indicate spatial extent and/or the profile of intrinsic object axes. Extent refers to the stretch of space an object occupies and is often expressed by using parts of the hands or arms to indicate endpoints; cf. Table 10.1 (a). The term profile refers to the course of the object's boundary and usually involves some kind of motion (b–h). Dimensional gestures often depict abstract one- or two-dimensional characterizations of the three-dimensional object (*dimensional underspecification*). Gestures (a)–(c) in Table 10.1 are one-dimensional, that is, depict an extent along one 'line'. Gesture (a) expresses extent as the space between the hands, while (b) and (c) additionally indicate the profile of this one-dimensional extent via movement. Gestures (d)–(f) are two-dimensional. All of them indicate the round profile of the reference object and the

TABLE 10.1. *A subset of the most frequent gesture kinds represented by prototypes*

Prototype	Description
	(a) flat hands, palms facing each other; indicates extent between left and right hand
	(b) extended index finger; fingertip moving straight; orientation perpendicular to movement; indicates extent
	(c) extended index finger; hand moving along index direction which indicates the extent; used mainly to depict an *interior* path, i.e. holes
	(d) extended index finger; fingertip describes a circular trajectory; fingertip movement indicates extent and profile
	(e) rounded C-hand-shapes; circle open or closed; posture indicates extent and round profile
	(f) flat hands, fingers aligned; hands perform semi-circular mirrored movements, palms facing towards the centre of the circle; indicates extent and round profile
	(g) hand is moving straight, perpendicular to the aperture; hand-shape indicates extent and round profile in two dimensions, movement adds another dimension
	(h) hands form an open or closed circle; hands moving downwards; hand-shape indicates extent and round profile in two dimensions, movement adds another dimension
	(i) flat hand, fingers aligned; hand moves into a direction parallel to the plane of the palm; movement and hand surface indicate a face of the object
	(j) flat hand as a placeholder; indicates orientation of an object in space

extent (i.e. the diameter) either by hand-shape or by movement. Gestures (g) and (h) are three-dimensional. In both cases a two-dimensional profile created via a distinct hand-shape is extruded by a linear motion resulting in the depiction of a cylindrical shape. A detailed analysis of the usage of gesture kinds for the depiction of certain objects or parts shows that an object's relative sizes (i.e. length: width: height ratio) partly determine the kind of gesture employed for a depiction.[2] A linear movement as in (b), (g), and (h) usually indicates a dominant extent (e.g. the length axis of the bar), whereas it is not used for an object or part lacking a dominant axis such as the cube. In contrast, two-handed, delimiting gestures as in (a) are used similarly for dominant extents, but also for the equally-sized extents of the compact cube. Generally, hand movement and two-handed delimitation are employed for the dominant extents, while hand-shapes are typically used to display subordinate extents. Besides relative size, we also compared absolute object and part sizes (as they appear on the screen) to the size of gestures if this gesture kind was employed for a depiction. We found a great variance of gesturally indicated sizes for an object with constant size. From this we conclude that it is relative rather than absolute size which is reflected in a person's dimensional gesture.

10.2.1.2 *Surface property gestures*

While dimensional gestures refer to the whole shape in terms of extent and profile, surface property gestures depict certain elements or features of an object's surface without reference to the whole object. Prototype (i) in Table 10.1 is an example of this type: the flat, moving hand indicates a particular planar side of the object without referring to the whole.

10.2.1.3 *Placeholder gestures*

These gestures are characterized by a body part representing the object itself. Spatial position and/or orientation properties are directly conveyed by the appropriate orientation of the body part in space. The realizations thus consist only of one-handed gestures with a distinct hand- or arm-configuration taking the approximate shape of the object. Prototype (j) is an example of a placeholder gesture. The whole hand stands for a longish, flat object and indicates its orientation in space.

10.2.1.4 *Spatial relation gestures*

This last gesture type indicates the relative position and/or orientation of two object parts using one hand for each. Thus, spatial relation gestures are always two-handed and usually asymmetrical. They may also consist of a combination of two individual gestures from the aforementioned types.

[2] Cf. Sowa (2006a) for a discussion of the effects of object sizes and proportions on shape-related iconic gestures.

Dimensional gestures account for 86% of all gestures, shape property gestures for 6%, and placeholder and spatial relation gestures for 2% each. Given the dominance of dimensional gestures in the corpus, it seems appropriate to consider the semantic features they express, namely extent and profile, as basic features for a representation of gesture content. Dimensional underspecification further implicates a consideration of extents and profiles independently for each spatial dimension. A semantic representation should reflect this underspecification, that is, it should be possible to specify just one dimension or object axis and to make no assumptions about the remaining dimensions.

10.2.2 Object decomposition

Some of the stimulus objects are easily decomposable into parts, for instance the screws can be composed into shank, head, and slot. Subjects usually realized this canonical object structure in their descriptions. Two object classes that apparently affect the way subjects describe the whole object can be distinguished. When the object's main body was a basic 3D geometry, like the bars and the cube, it was depicted in a gesture. For compositional objects like screws that consist of two almost equally sized parts, fewer gestures for the whole body were employed. In no case would a gesture depict the complex object shape all at once—for instance, drawing an outline of the screw as a T-shaped object. However, several subjects did depict the whole screw in an abstract way reducing it to its main extent.

10.2.3 The spatial organization of successive gestures

Gestural expressions have the potential to organize 'referents' in space and to build larger structures of meaning (Emmorey, Tversky, and Taylor, 2000; Enfield, 2004). Consistency between successive *verbal* expressions has already been found, as discussed in Vorwerg (this volume). Gestural expressions are also spatially cohesive in the sense that successive gestures often employ space in a consistent way (McNeill, 1992). Examples of spatial organization can be found in the corpus data.

Consider, for example, the gestures accompanying the description of the short bar (Figure 10.2). The subject first anchors the bar in space using a two-handed symmetrical gesture indicating its longitudinal extent. The left (non-dominant) hand is held in this position, while the right (dominant) hand indicates the position and shape of the holes with three successive strokes (meaningful phases). With the initial two-handed gesture, an imagistic context introducing the main object is set up in space. The validity of the context is explicitly bound to a visible feature, namely the left hand which keeps the position and shape of the initial gesture. This kind of organization we call *explicit spatial cohesion*.

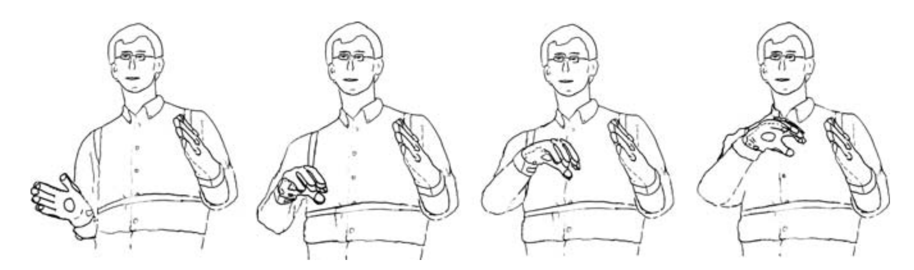


Fig. 10.2. Explicit spatial cohesion via a two-handed gesture. Left hand is held in position.

Conversely, there is *implicit spatial cohesion* whenever the spatial relation of successive gestures reflects the relation of the reference objects, but without any visible feature indicating cohesion. Figure 10.3 illustrates examples in which the spatial arrangement of successive gestures coincides with the spatial relation of the objects they refer to. Spatial cohesion can bind together several semantic entities (extents, profiles) either of a single object or part, or of two or more different objects or object parts. In Figure 10.3b we see an example of the former case, called *intra-object cohesion*. The dominant dimensions of the bar (its length and width) are displayed successively with two-handed gestures (indicating parallel lines) providing a two-dimensional specification of a single object (the bar). An example, of the latter case, *inter-object cohesion*, is depicted in Figure 10.3c. Three cohesive gestures successively indicate different parts of the screw: the shank (lower vertical line), the head (upper vertical line), and the slot (horizontal arrow).

10.2.4 Parts of speech associated with iconic gestures

Table 10.2 shows the frequency distributions of the parts of speech among the lexical affiliates (1st column, for n = 478 affiliates), their base frequencies in a representative sample of the whole corpus (2nd column), and the relative frequency of parts of speech among affiliates with respect to their base frequency (1st column divided by 2nd column). It is evident that nouns and in particular adjectives are overrepresented among the affiliates (relative frequency > 1.0), while the other classes are underrepresented (relative frequency < 1.0).

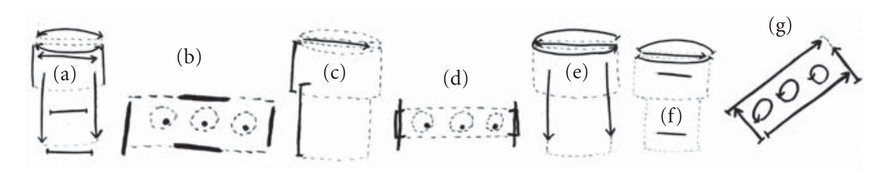


Fig. 10.3. Implicit spatial cohesion. Solid lines indicate gesture locations (arrows stand for movement in dynamic gestures); dotted lines show the reference object.

Table 10.2. *Frequency of the word classes among the affiliates and relative to the whole corpus*

	Affiliates (%)	Corpus sample (%)	Relative (affiliates/corpus sample)
Nouns	42.9	27.2	1.58
Adjectives	29.5	6.2	4.79
Verbs	4.0	15.4	0.26
Prepositions	5.2	8.3	0.63
Adverbs	14.2	21.8	0.65
Determiners	4.2	15.6	0.27
Interjections	0.0	5.5	0.00

A semantic analysis of the affiliated nouns shows that they include references to 3-D shapes such as *cylinder*, 2- or 1-D part references such as *side*, *face*, or *corner*, usually expressed after the introduction of the whole object in the discourse context, and references to object dimensions such as *length* or *diameter*. Affiliated adjectives similarly include 3-D descriptors such as *cylindrical*, 2-D expressions such as *round* or *six-sided*, and dimensional adjectives like *long* or *flat*. Furthermore, there are adjectives such as *flattened* or *slit* describing shape properties (modifications) of base objects, and other adjectives not directly related to shape but to object orientation and position. Most of these verbal affiliates express aspects of object extent, as in the case of dimensional adjectives, or aspects of extent combined with profile (boundary) properties as in 3-D nouns and adjectives. This shows that affiliates could refer to all spatial dimensions, or specify just two dimensions or one dimension of the object.

10.3 A Unified Shape Representation for Multimodal Signals

Taken together, the corpus evaluation revealed three important factors to consider in a semantic representation of shape-related gestural and verbal expressions. First, extent and profile are directly expressed in (dimensional) gestures as well as in accompanying adjectives and nouns and could be considered two basic semantic factors. Second, these elements are not expressed in isolation, but structurally organized in a spatially cohesive context. Third, a semantic representation should thus reflect the spatial arrangement of successive gestures. In the following, a shape-representation model that covers these factors is described. It extends an earlier approach which models the two factors of extent and (partly) profile information in gestures, but which has not included structured spatial organization of gesture and accompanying speech reflecting this factor (Sowa and Wachsmuth, 2002). A more detailed description of the formal structure can be found in Sowa (2006a, 2006b).

Models for shape representation that may inform the multimodal modelling approach can be found in different research disciplines including visual cognition, spatial reasoning, and linguistic modelling. Shape representations are usually divided into boundary- and interior-based approaches. The former primarily describe 2-D surfaces while the latter represent 3-D volumes. Interior-based approaches appear more relevant for the task because object shape is primarily a 3-D property. A further distinction can be drawn between quantitative and qualitative representations. Purely quantitative approaches are usually employed in computer graphics and geometric modelling where precision is needed (Mortenson, 1997). Yet, as the present study suggests, spatial information in gesture (and speech) is often abstract, qualitative, and underspecified. Precise approaches lack the capacity for abstraction and so their applicability is limited. Cohn and Hazarika (2001) provide a summary of representations for qualitative spatial reasoning.

One qualitative, interior-based method is to use volume primitives for shape approximation. This approach is exemplified by the geon model suggested by Biederman (1987) and the 3-D model by Marr and Nishihara (1978), which also introduces different levels of shape abstraction. However, geons and other volume primitives do not allow dimensional underspecification because they are inherently defined in 3-D. A one-dimensional gesture specifying a single extent could not be adequately represented. A suitable approach for the definition of the principal extent(s) of objects is provided by Lang (1989) within a semantic theory for dimensional adjectives. Lang defines representations called object schemata describing the basic gestalt properties of objects. Similarly, Clementini and Di Felice (1997) suggest properties for basic gestalt descriptions. However, none of these models fulfils all requirements that arise from the corpus analysis. Therefore, a new representation, called the Imagistic Description Tree (IDT), is proposed in the sections to follow, unifying the benefits of the model types above.

10.3.1 Modelling extent properties

For the modelling of extent properties we adopt the idea of an object schema as proposed by Lang (1989). Each object is described by a collection of up to three axes which represent the object's extents. An axis may cover one, two, or three spatial dimensions. A schema for a cylinder, for instance, would contain two axes. The first axis describes its height and is associated with one dimension. The second axis is associated with the remaining two (indistinguishable—due to rotational symmetry) dimensions. The 'object schema' representation is appropriate as a model for object descriptions with dimensional gestures and adjectives/nouns because it singles out the axes and their relations and thus allows dimensional underspecification.

Using different combinations of axes in an object schema, several basic object types with specific spatial characteristics can be represented as illustrated in

Table 10.3. *Representation of basic object types with object schemata*

Object schema	Prototype
$\{(1,\{\varnothing\},\perp), (1,\{\varnothing\},\perp), (1,\{\varnothing\},\perp)\}$	
$\{(1,\{max\},\perp), (1,\{\varnothing\},\perp), (1,\{\varnothing\},\perp)\}$	
$\{(1,\{\varnothing\},\perp), (1,\{\varnothing\},\perp), (1,\{sub\},\perp)\}$	
$\{(1,\{max\},\perp), (1,\{\varnothing\},\perp), (1,\{sub\},\perp)\}$	
$\{(1,\{\varnothing\},\perp), (2,\{\varnothing\},\perp)\}$	
$\{(1,\{max\},\perp), (2,\{sub\},\perp)\}$	
$\{(2,\{\varnothing\},\perp), (1,\{sub\},\perp)\}$	
$\{(3,\{\varnothing\},\perp)\}$	
$\{(1,\{\varnothing\},\perp), (1,\{\varnothing\},\perp)\}$	
$\{(1,\{max\},\perp), (1,\{\varnothing\},\perp)\}$	
$\{(2,\{\varnothing\},\perp)\}$	
$\{(1,\{max\},\perp)\}$	

Table 10.3. The first schema describes an object with three discernible axes of almost equal size. The typical instance, or prototype, for an object with these characteristics would be a cube. The first eight schemata in the table specify shape in three dimensions (illustrated with 3-D prototypes), while the last four show cases of dimensional underspecification. In Table 10.3 and in the following we use the following formal notation: curly brackets {...} delimit an *object schema*. The bracketed elements describe the *object axes*. An object axis has three properties: (1) a number (1, 2, or 3) describing how many canonical dimensions it is associated with; (2) a set of qualitative properties called *dimensional assignment values (DAVs)*; and (3) a measure for the numerical extent of the axis (e.g. in centimetres). We use triplets in round brackets (...) to indicate these properties. If a particular axis within a schema is labelled with the DAV *max*, it is the one with the largest numerical extent which corresponds to the length of the object. The DAV *sub* stands for substance and expresses minimality of the extent as compared to the other axes. It corresponds to object thickness. The unspecified DAV ∅ stands for an axis which is not significantly different in extent from the other axes in a schema. We use the symbol ⊥ to indicate that a numerical extent of an axis is not specified.

Here are two examples of how the spatial properties inherent in dimensional adjectives, nouns, and gestures can be represented using this model. Consider the adjective *longish*: its conceptualization in terms of an object schema would be $\{(1, \{max\}, \perp)\}$. This means that a longish object is characterized by an object

schema containing at least one axis which covers a single dimension and which is quantitatively most extended. Similarly, dimensional gestures can be semantically encoded using object schemata. Consider gesture prototype (h) in Table 10.1. The hands symmetrically form a round shape which is combined with a downward motion. Assume further that the extent of the motion is 40 cm, and the extent (diameter) of the circle formed by both hands is 20 cm. Both the movement component and the hand-shape are assumed to indicate spatial extent on the highest level of abstraction. The corresponding semantic encoding would thus be a schema containing two axes, that is, a one-dimensional axis representing the movement extent, and a two-dimensional axis representing the extent of the hand-shape, i.e. {(1, {max}, 40.0), (2, {sub}, 20.0)}.

10.3.2 Modelling profile properties

While extent properties refer to the basic proportions of an object, profile features provide additional information on the object's boundary. We adopt three general properties (symmetry, size, and edge) from the geon model here, although with some modifications. The *symmetry property* expresses regularities of the boundary with respect to one axis or a symmetric relation between two axes. The *size property* reflects the change of the extent of an axis when moving along another axis. The *edge property* determines whether an object's boundary consists of straight segments that form sharp corners, or of curvy, smooth edges. Profile properties are defined by a profile vector containing symmetry, size, and edge properties for each object axis or pair of axes. Considering gesture prototype (h) in Table 10.1 again, it is possible to infer profile properties of the object expressed in the gesture in addition to the basic extent information encoded in {(1, {max}, 40.0), (2, {sub}, 20.0)}. The fact that the hand-shape does not change during the downward movement indicates a constant extent (diameter) along the movement axis. This can be captured with a profile vector containing a size entry for 'constancy' of the second axis when moving along the first. A combination of two static gestures, for example (e) + (a) in Table 10.1, would not provide such profile information. Another example for the use of profile properties is given in the following section.

10.3.3 Modelling structure by an IDT

Object schemata are the building blocks of the IDT. They provide a description of an object's overall proportions and its major profile properties, but do not model structure and spatial relations. To this end, schemata can be arranged in a tree similar to the hierarchical structure used in the Marr and Nishihara (1978) model.

Structural aspects are represented in *imagistic descriptions*. An imagistic description for an object consists of a set of imagistic descriptions of its parts, an object schema defining its overall proportions, a spatial anchor flag, and a transformation matrix. This recursive definition provides a tree-like structure.

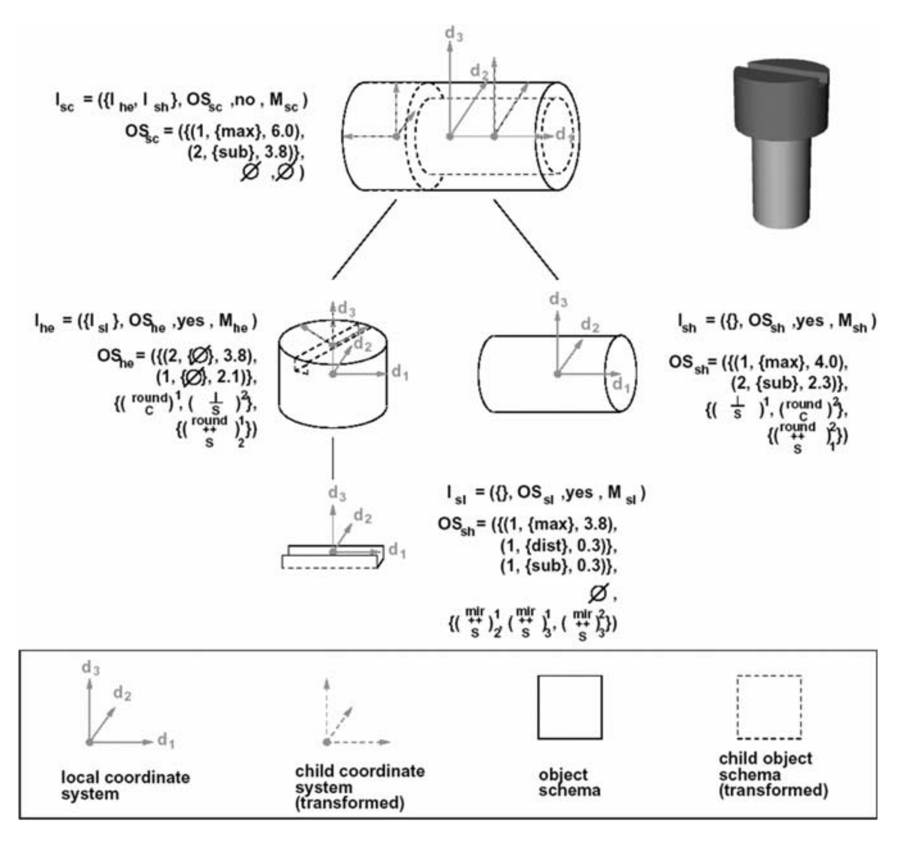


FIG. 10.4. Example of an IDT representation for a stylized screw.

The parts described are imagistic descriptions which could themselves contain further parts. The number of children is arbitrary. The spatial anchor flag signals whether the description is spatially anchored in a parent coordinate system. If its value is *yes*, the transformation matrix defines the position, orientation, and size of the object or part in relation to the parent description. An imagistic description of a perceived gesture, for instance, is spatially anchored because the gesture is performed in space and can be assigned spatial coordinates, while an imagistic description of an adjective is not spatially anchored. The complete tree describing an object including all parts, parts of parts, etc. is called an Imagistic Description Tree (IDT).

Figure 10.4 shows an example of an IDT model for the screw. The part hierarchy modelled by the three layers of the tree follows its perceptually salient decomposition. The top-level node I_{sc} represents the whole screw and has two child nodes modelling the parts, I_{he} for the head and I_{sh} for the shank. The head has another child node I_{sl} representing the slot.

Without providing all formal details of the IDT definition (Sowa, 2006a, 2006b), a closer look at node I_{he} representing the head will suffice to illustrate the model. The imagistic description I_{he} defines the slot representation I_{sl} as the only part. OS_{he} is the object schema that defines the basic proportions (axes) of the head. This contains two axes: the first covers two dimensions ($d1$, $d2$) and represents the 'diameter' with a numerical extent of 3.8 units; the second covers one dimension ($d3$) and represents the 'height' of the cylinder, which is 2.1 units. Since there is neither a perceptually dominant axis corresponding to 'length', nor a subordinated one corresponding to 'thickness', both axes are qualitatively described by the unspecified DAV Ø. The object schema definition is further augmented by profile vectors. It contains, for instance, the entry *(round, C)* for the first axis, where *round* is a symmetry property and expresses rotational symmetry of the axis, and C describes the curved boundary.

10.4 **Using the IDT in a Prototype System**

The IDT model forms the conceptual basis for representing shape-related information acquired via gesture and speech for usage in an operational gesture understanding system. The applicability of the IDT representation and a gesture and speech processing model have been tested with a prototype system. Gesture (motion) data is captured via data-gloves and motion trackers. The system is able to recognize and to conceptualize shape-related gestures and verbal expressions and to determine target objects which most closely match the input. To give a rough idea, the process of interpretation is outlined in Figure 10.5. Gesture and speech are perceived and segmented. The results of the segmentation process are uninterpreted surface descriptions of single words and gestures. For gestures, this surface description consists of a collection of spatiotemporal features.

Two decoders, one for each modality, convert the surface descriptions into elements of an IDT representation. The word decoder retrieves an adjective's or noun's semantic representation from a lexicon in terms of a complete IDT. The

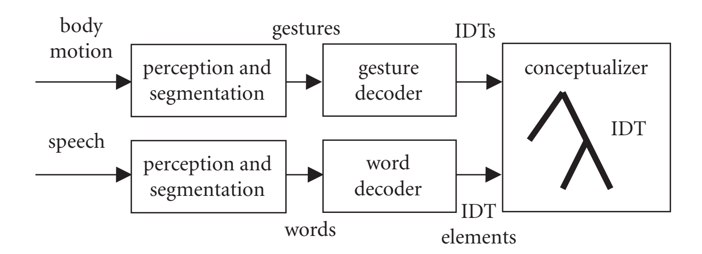


Fig. 10.5. Interpretation process.

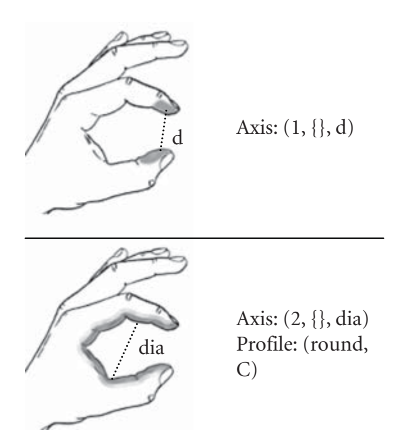


Fig. 10.6. Two different semantic interpretations of the 'C'-hand-shape in terms of IDT elements.

gesture decoder analyses the spatiotemporal features and transforms them into a set of object axis descriptions according to the form–meaning relations observed in the study. Since gesture and speech can be ambiguous, both decoders may output a set of alternative interpretations.

Figure 10.6 illustrates the decoding of a C-shape hand gesture. Subjects used it in two different ways (hand regions marked grey): to indicate the extent between the thumb and index fingertips and to depict a round profile with the curvature of the fingers. The former interpretation is represented by a 1-D object axis *(1, {}, d)*, while the semantics of the latter is described by a 2-D object axis *(2, {}, dia)* with additional boundary information *(round, C)*. Which one of the two interpretations is correct cannot be determined at this stage without further contextual information. Thus, both of them are forwarded to the next processing stage. The subsequent processing stage, called *conceptualizer* in rough accordance with the speech production and comprehension model suggested by Levelt (1989), maintains a spatial context model in the form of a dedicated IDT. This model can be considered the system's 'spatial imagination'. In the conceptualizer, incoming interpretations from the decoders are unified with the current model.

Integration of IDTs from verbal information is formally accomplished via a unification procedure that merges two compatible IDTs into a single one. Object axes resulting from gesture interpretation are inserted into the existing IDT. That way, successive gestures and words are integrated step by step, producing a unified spatial representation of an object description. Alternative interpretations may be ruled out during unification due to incompatibilities. Eventually, only one unified interpretation remains.

10.5 Discussion

While codified gestures with a fixed, culture-dependent meaning and pointing gestures have been described to some extent in the literature, there is not much work on the morphology and semantics of iconic gestures in specific domains. Exceptions are the early work by McNeill and Levy (1982), who first examined verbal and gestural representations used to depict cartoon narrations, and, more recently, Kopp, Tepper, and Cassell (2004), who examined iconic gestures in route descriptions for gesture synthesis in embodied conversational agents (see also Striegnitz *et al.*, this volume). A comprehensive study of the use of gestures for product design processes including shape description was conducted by Hummels (2000). In contrast to the work presented here, her study focuses not on coverbal but autonomous gestures performed independently of speech.

Computational models for the comprehension of iconic gestures and semantic fusion with speech are rare. Most work has instead focused on pointing gestures or symbolic gestures regarding gesture comprehension as a mere pattern classification problem. Seminal work on the understanding of iconic gestures for object placement and movement descriptions was done by Koons, Sparrell, and Thorisson (1993). Yet their approach is focused on applying spatial transformations depicted with iconic gestures ('place the box here', 'move it like this') to objects, but not on the integration of verbal and gestural information.

Our work addresses the lack of a semantic foundation for the integration of iconic gestures and speech in a specific domain. Extending the approach to multimodal dialogue systems, the IDT model could serve as a part of the spatial discourse context shared between a human and a computer system embodied in a virtual agent or robot. This would enable both interlocutors to fill space with meaning.

10.6 Conclusion

This chapter has addressed the meaning of shape-related iconic gestures. It has considered how such meanings can be accessed and modelled, as well as how they can be unified with the semantics of shape-related verbal expressions. Based on a comprehensive corpus of speech-gesture shape descriptions acquired from an empirical study, we have proposed the Imagistic Description Tree (IDT) as a representation for the semantics of multimodal shape-related expressions, and outlined its application in a gesture understanding system. The IDT models object extent, profile, and structure as the salient semantic elements contained in gesture and speech. The IDT representation is an important step towards capturing the meaning of iconic gestures in formal terms and making possible their computational treatment together with speech. An application has been outlined that can algorithmically generate interpretive operational shape descriptions from gesture and speech input modalities.

11

Knowledge Representation for Generating Locating Gestures in Route Directions

KRISTINA STRIEGNITZ, PAUL TEPPER, ANDREW LOVETT, and JUSTINE CASSELL

11.1 Introduction

When giving route directions, humans may use gestures for a variety of purposes, such as indicating turns and movement direction, to describe the location of landmarks, and to depict their shape (see also Sowa and Wachsmuth, this volume). In previous work (Kopp, Tepper, Ferriman, Striegnitz, and Cassell, 2007), we have studied how gestures are used to describe the *shape* of landmarks and how such gestures can be generated in an embodied conversational agent (ECA). In this chapter, we look at the way humans use gesture to indicate the *location* of landmarks. Emmorey, Tversky, and Taylor (Emmorey, Tversky, and Taylor, 2000; Taylor and Tversky, 1996) have found that people alternate between different perspectives when giving directions. We examine the use of these different perspectives in our data (Section 11.2). Next, we formulate requirements on knowledge representation for generating such gestures in an ECA (Section 11.3), and we propose a way of implementing these requirements (Section 11.4). We then sketch how this information is used in a direction-giving ECA (Section 11.5). Finally, Section 11.6 relates our results to previous work before we conclude in Section 11.7.

11.2 Gestures in Direction-Giving Dialogues

11.2.1 Data

The observations described in this chapter are based on videos of people giving directions across Northwestern University's campus to another person who (they believe) is unfamiliar with the campus. In addition to transcribing the speech, we have identified and coded gestures referring to landmarks, annotated them

TABLE 11.1. *Distribution of statement utterance units over statement type*

statement type	# of utterance units
reorient	32
reorient+lm	24
move	51
move+lm	119
lm	367
dir	3
	597

with their referents (a basic name for what they seem to depict) and information about the perspective used (as described below). Utterances have, furthermore, been marked for the dialogue moves that they accomplish, using a coding scheme that was inspired by the DAMSL coding scheme (Allen and Core, 1997) and by the scheme for classifying instructions in route directions introduced by Denis (1997). The scheme is also similar to the one used by Muller and Prévot (this volume) to annotate French direction-giving dialogues with dialogue moves.

We coded five direction-giving dialogues, which altogether consist of 753 utterance units by the person giving the directions and 234 utterance units by the person receiving the directions. We are interested for the purposes of the present chapter in the direction giver's language and will, therefore, concentrate on these contributions to the dialogue. Utterance units are annotated along five different dimensions. First, they are classified with respect to their communicative status and information level. 640 of the direction giver's utterance units are interpretable and directly pertain to the task. All others were either abandoned, otherwise uninterpretable, or meta-communications about the task or conversation.

The second dimension marks utterance units that make assertions contributing to the route description as statements. We distinguish six types of statements: instructions to reorient, or to reorient with respect to a landmark (labelled as reorient and reorient+lm, respectively), instructions to move, or to move with respect to a landmark (move/move+lm), statements that mention a landmark without an instruction to reorient or move (lm), and statements describing cardinal directions (dir), such as '*north is that way*'. 597 of the 640 utterance units by the direction-giver (that is, 93%) are statements. Table 11.1 shows the distribution of utterance units over statement types.

Our third and fourth dimensions look at queries and responses marking clarification questions (Q-clarif), requests for feedback (Q-feedback), and other requests for information (Q-other), and answers to clarification questions (A-clarif), back-channel feedback (A-ack), and other answers (A-other). 18 of the

direction-giver's utterances (3%) are queries and 185 (29%) are responses. 172 of the responses are answers to clarification questions and 13 are back-channel feedback. Note that the statement, query, and response dimensions are not mutually exclusive. For example, many statements (158) are part of a response. Therefore, the totals for statement-, query-, and response-type utterance units do not add up to 640 or 100%.

Finally, we mark utterance units that belong to an elaboration on a landmark or action (elab), such as the second utterance in 'The Allen Center is to your left. It's really big', or that are part of a redescription of a route segment that has previously been introduced and described (repeat). In our data, 227 utterance units are annotated as elaborations and 75 as part of a redescription. All of them are statements.

11.2.2 Perspective of locating gestures in direction-giving dialogues

The literature on route descriptions discusses two perspectives that people use for describing space along the route (Taylor and Tversky, 1996). In *route perspective*, landmarks are described in the frame of reference of a person walking the route. In contrast, the *survey perspective* is like a bird's-eye view. Buildings are described relative to each other or to an absolute frame of reference (for example, cardinal directions). These two different perspectives are also reflected in the gestures that accompany speech (Emmorey, Tversky, and Taylor, 2000), and we find examples of both perspectives in our data. We also find gestures that do not fall into these two categories. First, we find gestures that seem to be purely shape-depicting, and which do not refer to the location of the referent landmark at all. Second, we find gestures which locate the object with respect to the speaker's actual position and orientation.

Figure 11.1 shows an example of a gesture where the speaker has taken on the perspective of the person following the route (the route perspective). He speaks and gestures as if he has the position and orientation that an imaginary direction-follower would have at this point along the route. Therefore, the location of his gesture (to the left of his body) corresponds to the location of the landmark relative to the location and orientation of the imaginary direction-follower. This perspective is by far the most common in our data (54.2% of all gestures referring to landmarks).

Another way in which people use their hands and the space around their bodies is to lay out virtual maps using a bird's-eye view, as shown in Figure 11.3. Map gestures are unique in that, after one gesture is made, the hand is held in place, while the next location is depicted relative to the first by placing the other hand relative to the position of the first. As Figure 11.3 illustrates, the right hand representing University Hall is the anchor, held in exactly the same position throughout the three-gesture sequence, while the locations of Kresge and Harris Hall are shown relative to it. Kresge is shown using an almost identical gesture, a



Fig. 11.1. 'On your left once you hit this parking lot [is the Allen Center]'.

flat hand shape facing downwards, placing the building with respect to University. This probably indicates a survey perspective for these two gestures. Harris is not actually placed in the same way; rather it is pointed to in a kind of deictic gesture that assumes the route perspective, or the perspective of the imaginary direction-follower. This mixed-perspective interpretation is supported by the accompanying language, which serves to place the first two landmarks, University and Kresge, and indicates that the third, Harris, is not placed on the left or the right of the follower but 'straight ahead' of the follower. Overall, the virtual map is oriented in the same way, such that it matches up with the direction a person walking the route would be facing. We found that 16.3% of the landmark-depicting gestures in our data are survey-perspective map gestures.

It is important to note that gestures referring to landmarks do not necessarily have a locating function. For example, after having located the Allen Center to the left of the direction-follower, the speaker in Figure 11.1 continues by saying *and it's really big*. He accompanies this elaboration with the gesture shown in Figure 11.2, which refers to the landmark's shape by indicating its horizontal extent. This gesture does not locate the landmark to the left, which would be its position with respect to the point of view assumed for the previous utterance. Instead the gesture



FIG. 11.2. 'and [it's really big]'.

is carried out in front of the speaker's body. In our data, 15.8% of the gestures referring to landmarks are of this non-locating kind. However, often gestures are neither purely locating nor purely shape-depicting. For instance, the gesture used in Figure 11.1 seems to indicate the wall of the building being described as the shape of the hand is flat and vertically-oriented. It thus has a shape-depicting component in addition to its locating function. In this chapter, we are concerned with the locating function of gesture and will not address the issue of how to determine which shape features to depict and how to depict them (but see Kopp *et al.*, 2007, and Sowa and Wachsmuth, this volume, for more on these questions). Finally, gestures may be used to locate objects with respect to the speaker. That is, the speaker simply points to a real object. This type of gesture is extremely rare in our data (only 1.9% of all gestures referring to landmarks fall in this class). Table 11.2 shows the distribution of perspective among gestures referring to landmarks in our set of direction-giving dialogues.

11.2.3 Perspective and dialogue structure

In order to generate locating gestures with different perspectives, we must address the following question: When are the different perspectives used? As the following results show, the use of these perspectives seems to be determined in part by the dialogue move that the speaker is trying to perform. In our data, most of



FIG. 11.3. '[University Hall] is on your right, [on the left is Kresge], and [then straight ahead is Harris]'.

Table 11.2. *Distribution of perspective among gestures referring to landmarks*

perspective	# of gestures	%
route perspective	185	53%
survey perspective	57	16%
non-locating	58	17%
locating wrt. speaker	7	2%
unclear/ambiguous	40	12%
	347	100%

the direction-giver's gestures referring to landmarks occur with utterance units marked as statements. In fact, *all* of the survey-perspective, route-perspective, and non-locating gestures, which are the gestures we are most interested in, co-occur with statements. Table 11.3 shows which statement types the different gesture perspectives co-occur with. Unsurprisingly, gestures of any perspective that are referring to landmarks co-occur with utterances that mention a landmark in the speech. (Recall that we are not looking at gestures depicting actions here.)

None of the gestures under consideration co-occur with queries, but some of them co-occur with statements that are also marked as an elaboration, as a redescription of previously explained route segments, or as a response to a clarification question (we do not have cases of co-occurrence with other response types). Tables 11.4–6 show the frequency with which gestures of the different perspectives co-occur with utterance units with these labels. Table 11.7 shows how often gestures of the different perspectives co-occur with plain statements, that is, statements which are not marked as a response, a query, an elaboration, or a redescription. The tables also show the percentage deviation for those frequencies, which

Table 11.3. *Distribution of gesture perspective over statement type*

type of statement	# of survey-perspective gestures	# of route-perspective gestures	# of non-locating gestures	# of speaker-perspective gestures	# of unclear/ unambiguous gestures
reorient	0	1	0	0	1
reorient+lm	1	4	1	0	0
move	0	0	0	0	0
move+lm	2	23	2	1	6
lm	54	157	55	5	33
dir	0	0	0	0	0
	57	185	58	6	40

Table 11.4. *Frequency of gesture perspective in answers to clarification questions*

	# of survey-perspective gestures		# of route-perspective gestures		# of non-locating gestures		# of speaker-perspective/ unclear/ unambiguous gestures		
statement is A-clarif	32	+110%	31	−38%	18	+16%	12	−5%	93
statement is not A-clarif	25	−40%	154	+14%	40	−6%	35	+2%	254
	57		185		58		47		347

measures how much the frequency differs from the frequency we would expect if gestures were equally likely to co-occur with utterance units of any dialogue function.

Survey-perspective gestures occur much more often than we would expect in answers to clarification questions and in redescriptions of route segments. They occur much less often than expected in plain statements. This indicates that speakers switch to survey perspective when they need to re-explain a portion of the route. It also fits findings of a previous study on direction-giving, which differed from our own in that the subjects could use a physical map (Cassell *et al.*, 2002). In that study, subjects only referred to the map if their purely verbally given directions were not sufficient.

In contrast, route-perspective gestures occur more often than expected in plain statements and less often in statements marked as A-clarif, elab, or repeat. So, the route-perspective seems to be the default when gesturing about landmarks. Non-locating gestures, finally, occur much more often than expected in

Table 11.5. *Frequency of gesture perspective in elaborations*

	# of survey-perspective gestures		# of route-perspective gestures		# of non-locating gestures		# of speaker-perspective/ unclear/ unambiguous gestures		
statement is elab	22	−13%	60	−27%	51	+98%	21	+1%	154
statement is not elab	35	+10%	125	+22%	7	−78%	26	−1%	193
	57		185		58		47		347

Table 11.6. *Frequency of gesture perspective in re-descriptions*

	# of survey-perspective gestures		# of route-perspective gestures		# of non-locating gestures		# of speaker-perspective/ unclear/ unambiguous gestures		
statement is repeat	16	+144%	13	−39%	7	+5%	4	−26%	40
statement is not repeat	41	−19%	172	+5%	51	−1%	43	+3%	307
	57		185		58		47		347

elaborations and much less often in plain statements. They occur slightly more often than expected in answers to clarification questions. This can be explained as follows. After having introduced a landmark, probably using a gesture that locates the landmark, speakers give further information about the visual properties of the landmark, such as its shape or size. This is reflected in their gestures in which the locating component may be absent or deemphasized.

11.3 Requirements on Knowledge Representation

To generate any kind of route description, a map of the relevant area is needed. Minimally, the map must include the paths that can be taken, so that the system can calculate the route. Unlike direction-giving systems such as MapQuest, our system gives directions using landmarks to indicate reorientation points and other

Table 11.7. *Frequency of gesture perspective in plain statements*

	# of survey-perspective gestures		# of route-perspective gestures		# of non-locating gestures		# of speaker perspective/ unclear/ unambiguous gestures		
statement is plain	2	−90%	96	+48%	7	−66%	17	+3%	122
statement is not plain	55	+49%	89	−26%	51	+36%	30	−2%	225
	57		185		58		47		347

important points along the path. Therefore, our map representation has to include the landmarks located along these paths. As the data presented in Section 11.2 show, gestures referring to these landmarks may express different perspectives. The perspectives differ in whether or not and how relative location in the map representation is reflected in the placement of gestures in the gesture space. This requires information about the position and orientation of both the imaginary direction follower and the speaker as well as mechanisms for inferring spatial relations between entities in the map and mapping them to the speaker's gesture space.

For survey- and route-perspective gestures, we need to keep track of the position and orientation that a person following the route would have at each point of the description. And, in order to generate gestures which locate landmarks relative to the speaker, we need the position and orientation of the person or ECA giving the directions in the map. The system also requires mechanisms for inferring spatial relations between the entities in the representation. For example, the system needs to be able to infer the location of landmarks relative to paths, other landmarks, the speaker, and the direction-follower. This is necessary for deciding which landmarks to mention in the route description; landmarks that are mentioned at a specific point in the description should be visible to the direction-follower when he/she reaches the corresponding point of the route. In addition to these inference mechanisms, the system needs an appropriate mapping from positions in the map representation to positions in the gesture space in order to place both route- as well as survey-perspective gestures correctly in the gesture space. For example, the position of route-perspective gestures should reflect the relative location of the landmark with respect to the direction-follower, and the positions of the different gestures in a survey-perspective sequence should reflect the relative location of the landmarks to each other and to the direction-follower. Additionally, the discourse history has to contain information about the current location of the hands and which landmark they stand for, such that multimodal anaphoric expressions can refer back to these landmarks in later utterances.

Finally, landmarks and paths must be associated with semantic information. For instance, a description of a landmark could draw upon information about its name, type (building, lake, monument, etc.), size, colour, and shape. For paths, we may specify what type of path it is, a street, parking lot, courtyard, etc. This information is necessary for generating descriptions of landmarks together with gestures depicting their shape and/or size. In the next section, we propose a way of implementing the knowledge requirements formulated above in an ECA.

11.4 Locating Landmarks in Space

The basis for generating locating gestures is a map representation consisting of two interlinked components: (i) a graph, where edges represent the paths that can be

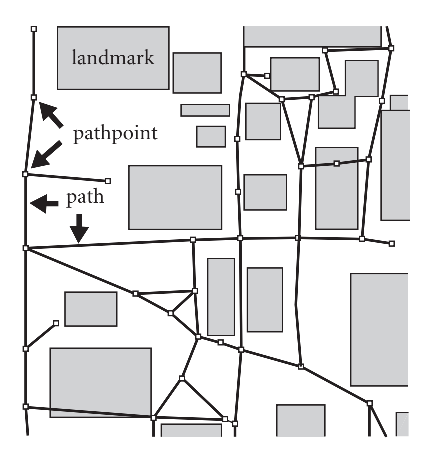


Fig. 11.4. Map representation showing path points, paths, and landmarks.

walked and nodes (path points) represent points on the map where the direction-follower may have to change his or her direction, and (ii) a set of landmarks. Landmarks are associated with areas and path points are associated with points in a common coordinate system (see Figure 11.4). In addition, path points can be linked to landmarks by qualitative relations specifying whether a path point is the entrance of a building or whether it is next to a landmark (Figure 11.5). Finally, landmarks and path points are associated with semantic information as described above (type of landmark, size, colour, shape, etc.). Note that Shi and Tenbrink (this volume) also present a discussion of the representation of spatial information for direction-giving and -following.

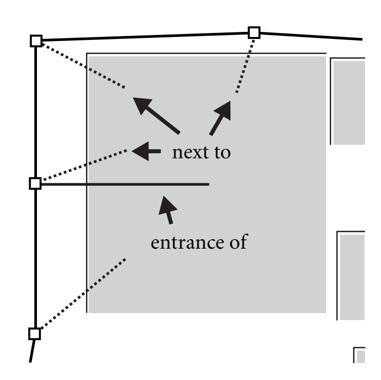


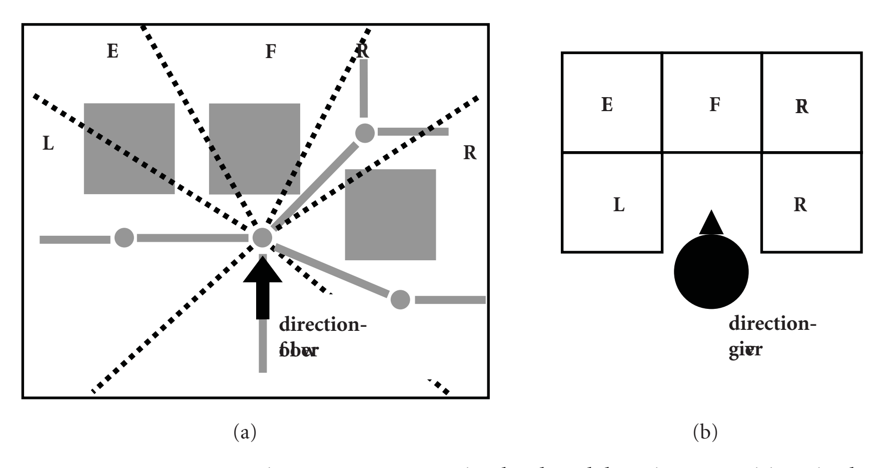
Fig. 11.5. A landmark with qualitative relations to path points.

11.4.1 Locating landmarks with respect to the direction-follower's and the speaker's perspective

When gestures are used to locate landmarks with respect to the *direction-follower*'s point of view, they depict the landmark at a location in the gesture space. This location corresponds to the location of the landmark relative to the position and orientation that the direction-follower would have in the world at that moment if he/she were walking the route. This holds whether it is a simple pointing gesture or a gesture that depicts some aspect of the landmark's shape, as in Figure 11.1. In order to generate such gestures, we need to keep track of the position and orientation of the direction-follower in the map representation. These values change continually over the course of the dialogue, as the description (and the imaginary direction-follower) progresses along the route.

Given a route between two points on the map graph, we can derive the direction-follower's orientation for each point along this route, based on the location of the previous point on that route. This allows us to calculate the angle at which landmarks are located with respect to the direction-follower's orientation, which can then be mapped to different positions in the speaker's gesture space. Since these gestures are normally only used to locate the landmark with respect to the direction-follower and do not represent relative location to other landmarks, we use a coarse mapping that maps ranges of angles to five different positions in the gesture space: left, right, front left, front right, and front (see Figure 11.6).

Gestures that locate objects with respect to the *speaker* can be generated using the same mechanisms, given that the location and orientation of the speaker are recorded within the map representation. Note that in our current application the



(a) (b)

FIG. 11.6. Route-perspective gestures—mapping landmark location to positions in the gesture space.

speaker is our ECA, which is part of a stationary information kiosk. The agent is displayed on a fixed screen, so its position and orientation remain the same over the course of an interaction.

11.4.2 Generating map gestures

In their simplest form, map gestures resemble the act of placing objects in the horizontal, tabletop plane in front of the speaker. While they can get more complicated than this, for example by also depicting information about the shape of the objects, here we will just consider this basic case of positioning objects. Neither are we currently modelling map gestures where route and survey perspective are mixed, as in the example in Figure 11.3. Each map gesture depicts a limited section of the map of the world. This section contains the target landmark and a number of other visible landmarks. We choose landmarks which either could easily be confused with the target or can help in distinguishing it. For example, if the target landmark is a building which is to the left of the direction-follower and there is another building which is also to the left or to the left and front, then the target could easily be confused with this second landmark based on their locations. Or if, for example, the target is a path turning only slightly left and there is another path continuing straight, these two paths can easily be confused and would both be included in a map gesture.

Once we have identified which landmarks to include in the map gesture, we compute the angles at which those landmarks are located with respect to the current position and orientation of the direction-follower in the map or, in the case of paths, the angle at which the path leaves this point. Those angles are then mapped to positions on an imagined circle which is centred slightly in front of the speaker's body in the tabletop plane. Positions on this circle are described in terms of the three-dimensional coordinate system representing the speaker's gesture space. Figures 11.7 and 11.8 show examples of this mapping. If we assume the target landmark in Figure 11.7(a) is building B, there is one building (building A) which could easily be confused with the target. So the relevant section of the map for the map gesture contains buildings A and B. Figure 11.7(b) shows the positions in the gesture space they are mapped to. Let us now assume that the target is the path labelled C in Figure 11.8(a). This path could easily be confused with path E, while building D can help to distinguish them. Figure 11.8(b) shows how paths C and E and building D get mapped to the gesture space.

The next step is to decide what gestures to use to indicate these locations and how to order them. We use a static gesture for buildings, which places a hand with a flat hand-shape and the palm pointing down at the point in the gesture space determined by the mapping. For paths we use a dynamic gesture which 'draws' a line from the centre of the imagined circle to the (end)point determined by the mapping. A pointing hand-shape (where the index finger is extended and all other fingers are curled up) is used. The order of the gestures making up the map gesture

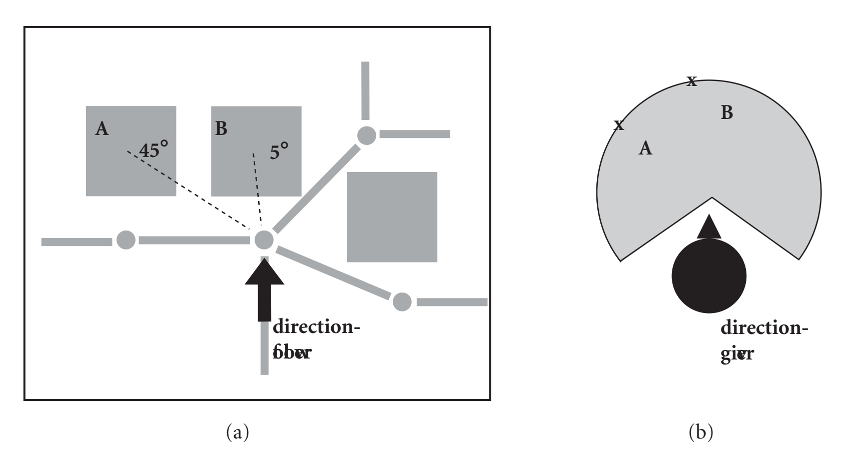


Fig. 11.7. Map gestures—mapping the location of buildings A and B to positions in the gesture space.

is determined as follows. Generally, the target is mentioned first and then all other landmarks going either clockwise or anticlockwise from the target. If the target is a path and some three-dimensional landmarks are involved in the map gesture, the three-dimensional landmarks are mentioned first, then the target, and then all other landmarks. Finally, we propose to store information linking the agent's hands to their locations and to the entities they represent in the dialogue context. This information needs to be updated appropriately as the relations between hands, locations, and landmarks change. This allows later utterances to make use of the information, for example, in order to generate appropriate multimodal

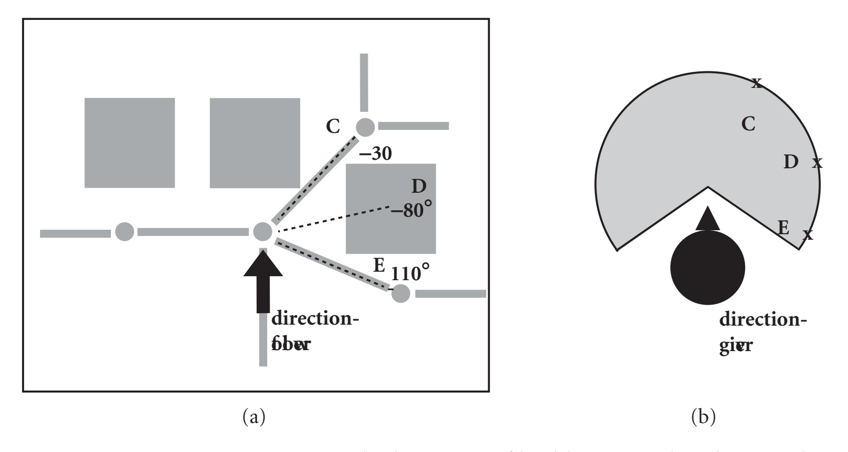


Fig. 11.8. Map gestures—mapping the locations of building D and paths C and E to positions in the gesture space.

anaphoric references to landmarks, where the ECA continues using the same hand and location to refer to the same landmark as long as the direction-follower's position and orientation remains stable.

11.5 Architecture of a Direction-Giving ECA

Now, we move on to describe the architecture of our ECA, called NUMACK, illustrated in Figure 11.9. First, we discuss the dialogue management module and its central data structure, the Information State. Next, we describe the content planning stage, which includes a route planner that employs a map representation specialized for gesture and natural language generation (see Section 11.4). The content planner also determines the perspective used in each gesture. Lastly, we give a brief description of the multimodal microplanner and surface realization components.

At the centre of the system is the Information State (Traum and Larsson, 2003). This is a data structure that keeps track of the dialogue history, the private knowledge of the system, the shared knowledge of user and system, and the current state of the system. In addition to this kind of information, which is commonly found in any Information State, we also use the Information State to store the output of the content planner, and to keep track of the point in the route the description has reached. We are still working on integrating the information necessary for producing anaphoric gestures as described in the previous section into the Information State. The Dialogue Move Engine determines how to integrate user dialogue moves into the Information State and chooses the moves of the system. We use Midiki, the MITRE Dialogue Kit (Burke *et al.*, 2003) in our implementation; this

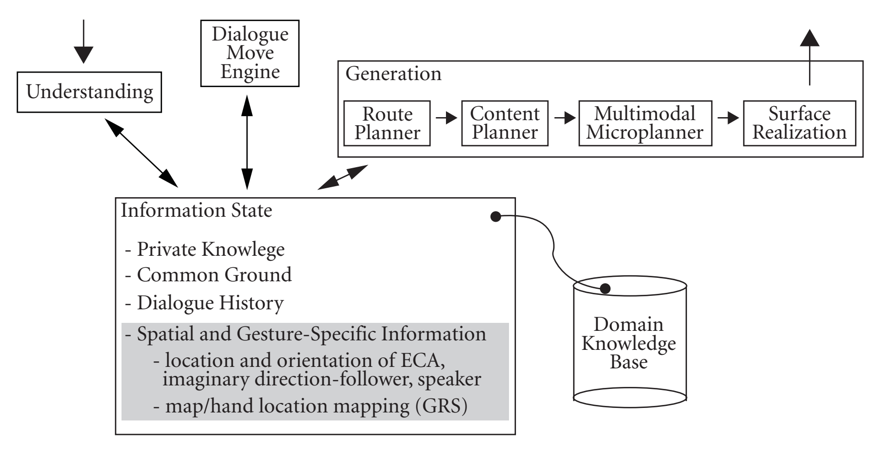


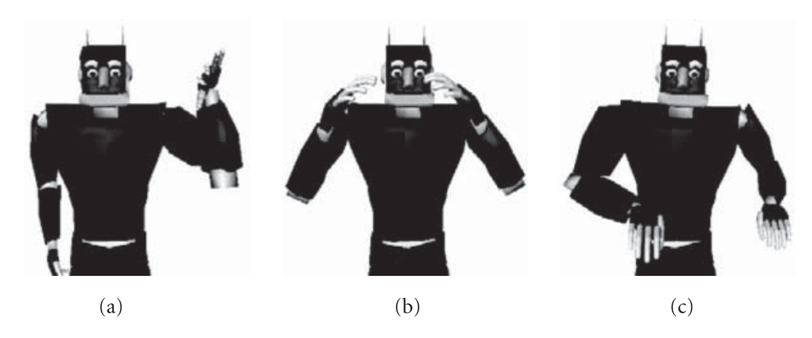
FIG. 11.9. Architecture of a direction-giving ECA.

provides a Dialogue Move Engine, lets us specify the rule system for the Dialogue Move Engine, and maintains the Information State for us.

Once the system has determined where the user wants to go and where he or she wants to leave from, the route planner calculates the shortest path between these two points. The map representation that the route planner currently works with has been coded by hand. Ultimately, we would like to automatically derive the necessary information from existing sources of geographic information. The output of the route planner is a sequence of path points and the task of the next step, which is content planning, is to map this to a sequence of preverbal messages, which can then be turned into multimodal utterances by the multimodal microplanner. More specifically, the content planner (i) chooses which path points to mention, (ii) decides which instruction types (that is, reorient, reorient+lm, move, move+lm, or lm) to use for describing each step in the route, (iii) selects landmarks that can be used to identify path points to the user, and then (iv) determines the semantic content of the expressions referring to those landmarks. In step (iv), the content planner chooses the properties of the landmark that need to be expressed either in language or in gesture to distinguish the landmark from its surroundings. It also determines the perspective that should be used with respect to gesture.

It is then in these last two steps that the data structures described in the previous sections come to bear. By default, the system assumes the route perspective. Figure 11.10(a) shows an example of a route-perspective gesture, which accompanies the words '*Pass the Allen Center on your left*'. Non-locating gestures are only used in elaborations on landmarks that do not mention the location of that landmark (for example, Figure 11.10(b): '*Dearborn Observatory is the building with the dome*'). As our system's capabilities to accept and react to clarification questions are still very limited, we only use map gestures for redescriptions of route segments. Such redescriptions are triggered if a reorientation occurs at a point where one or more turns are possible that can easily be confused with the target turn (cf. the situation in Figure 11.8(a)), or if the destination landmark can easily be confused with neighbouring landmarks (cf. the situation in Figure 11.7(a)). Figure 11.10(c) shows an example of such a map gesture. The accompanying speech is '*Annenberg Hall is here and the Seminary is here*' where the first occurrence of *here* refers to the position of the right hand and the second one to the left hand.

The output of the content planner specifies the structure of the route description and the semantic content that needs to be expressed by each utterance. It is stored in the Information State. Based on user feedback, the dialogue manager chooses when to send the next utterance specification to the microplanning and realization modules. The multimodal microplanner determines the form of the utterance, including the actual words as well as the form of the gestures and the coordination between language and gesture (Kopp *et al.*, 2007). Finally, the



(a) (b) (c)

Fig. 11.10. NUMACK, our ECA, producing (a) a route-perspective gesture, (b) a non-locating gesture, (c) a survey-perspective gesture.

surface realization component turns the utterance specification produced by the microplanner into speech and movements of the animated character on the screen (Kopp and Wachsmuth, 2004).

11.6 Related Work

Most literature on deictic gestures in multimodal interfaces concerns the *interpretation* of such gestures (see, for example, Bolt, 1980; Johnston and Bangalore, 2000). There are systems which *generate* deictic gestures, such as the COMIC system (Foster, 2004) and DFKI's PPP Persona (André, Rist, and Müller, 1998), but these systems only handle pointing gestures that point to objects presented on the screen. They are hence what we have called gestures that locate objects with respect to the speaker.

Another body of research that is relevant to our application is the existing work on generating natural language route descriptions. For example, Dale, Geldof, and Prost (2005) generate driving directions from GIS data. Look, Kottahachchi, Laddaga, and Shrobe (2005) produce walking directions, but concentrate on the representation of the information necessary for planning the route rather than the planning and realization of the natural language output. Habel (2003) concentrates on the architecture of a generation system for route directions, arguing for an incremental processing model. None of these systems models face-to-face dialogue; hence, none of them looks at generating the gestures that humans use when giving route directions.

More recently, Theune, Hofs, and van Kessel (2007) have described an ECA that generates route directions in a virtual environment. However, they do not generate words and gestures in an integrated way—the words are generated first, then gestures are added—and while their system has a mechanism for choosing

between different kinds of gestures they do not consider survey gestures and seem to mostly rely on pointing gestures from the direction-follower's point of view. As part of this research, Evers, Theune, and Karreman (2007) also investigate the effect that the orientation of the direction-giver with respect to the person receiving the directions has; however, whether the ECA is facing that person or is positioned to look in the same direction as that person was not found to influence the effectiveness of the directions. Nevertheless, the directions were perceived as more natural when the ECA is facing the user. As Figure 11.10 shows, NUMACK is facing the user.

11.7 Conclusions and Future Work

Previous work on human face-to-face dialogue has shown that speakers assume different perspectives when giving route directions (Taylor and Tversky, 1996). In particular, they use the route perspective, which refers to landmarks with respect to an imaginary direction-follower's point of view, and the survey perspective which locates landmarks using a bird's-eye view. Our data support this finding and also show that, in addition to route-perspective and survey-perspective gestures, people use non-locating gestures and gestures that locate landmarks with respect to the speaker's point of view. The distribution of these gestures is partly determined by the dialogue move of the utterance they occur in. Our goal is to model the different uses of locating gestures in a direction-giving ECA in order to produce route descriptions which are more natural and easier to understand. To the best of our knowledge, the issue of perspective in locating gestures has never been addressed with the aim of generating such gestures in a virtual agent.

This chapter has discussed the knowledge necessary for generating such gestures and we have proposed a way of representing this knowledge in an implemented system. More specifically, we have argued that we need a suitable map representation (representing not only the paths that can be walked on but also landmarks in relation to these paths as well as additional semantic information about properties of paths and landmarks) and that we have to be able to keep track of the position and orientation of entities in this map (that is, landmarks as well as the direction-follower and the speaker). This information is necessary for generating route-perspective and survey-perspective gestures as well as gestures that locate a landmark with respect to the speaker's point of view. In the case of map gestures, the position of the speaker's hands needs to be recorded and linked to landmarks, and this information needs to be appropriately updated as the discourse proceeds.

The proposal made in this chapter is implemented in a direction-giving ECA. We are currently preparing a study to evaluate the way this ECA uses gestures.

Furthermore, we are working on making the system more interactive. The main goal is to make it more effective by taking user feedback into account, but this will also allow us to further integrate our findings on how dialogue moves influence gesture perspective.

Acknowledgements

We would like to thank Magan Sethi and Laura Cristina Stoia for their help with implementation. This research was supported by a fellowship from the Postdoc Programme of the German Academic Exchange Service (DAAD).

12

Grounding Information in Route Explanation Dialogues

PHILIPPE MULLER and LAURENT PRÉVOT

12.1 Introduction

Speakers usually devote a lot of time in a conversation to negotiating the information they exchange. When that information concerns spatial or spatio-temporal objects, for instance route explanations, speakers have to agree (i) on a set of spatial landmarks and (ii) on their localization and, in the case of a route, how they can get from one point to another. Of particular interest to us is how positive feedback is used for spatial knowledge management and what the dynamics of settled information in a conversation (or Conversational Common Ground as defined in Clark, 1996) are. There have been numerous studies detailing the roles that assertions, questions, and answers can take in a conversation. Coding schemes for dialogue such as Dialogue Act Markup in Several Layers or DAMSL (Core and Allen, 1997; Carletta *et al.*, 1997), take great care in distinguishing these functions. Following the DAMSL (Core and Allen, 1997) terminology, dialogue acts are divided into those having mainly a forward-looking function and those having mainly a backward-looking function. Much attention has been given to the question/answer pair, its semantic and pragmatic interpretation (Ginzburg, 1996; Asher and Lascarides, 1998). Less emphasis has been placed on the role of all speech turns ensuring that information exchanged is properly interpreted (feedback). This is, however, crucial when many referents are possibly unknown to one of the participants in a dialogue, as is the case in route explanation dialogues. The important work of Traum (1994) has studied in some detail how these utterances play a role in deciding the status of information exchanged during a dialogue (mutually accepted or under discussion). He emphasizes that different levels of acknowledgement exist as described by Clark (1996) or Allwood (1995). We want to show here how distinguishing between such turns is central to the establishment of information, along with question/answer pairs, how they can be accounted for in a structural theory for representing dialogue, and how they interact with spatial information.

The role and influence of several discourse markers on acknowledgements in a French corpus of direction-giving dialogues have been investigated in order to study how speakers agree on specific locations that are part of an explanation. Spatial knowledge is thus seen here as a means of forming a precise analysis of the meaning of a certain type of dialogue act, while we can at the same time gain insights into the mental representation of routes by observing how this knowledge is shared.

Route explanations are a specific type of communication about spatial information that has been studied in psychology and psycholinguistics as very representative of human spatial cognition and the relation between spatial representation and language (Taylor and Tversky, 1992). This is probably because a route is a highly structured and complex piece of information, which is not straightforward to communicate or understand. It is common to see a route as a series of temporally ordered actions performed with respect to a set of spatial landmarks. Denis (1997) claims that actions are of two main types: change of orientation or plain progressions. In this respect, landmarks are seen as ways of locating actions and other relevant landmarks, or ways of checking the description.

When the explanation is given through an interaction between two speakers (as opposed to a monologue as in Denis, 1997), specific communicative issues appear: the receiver of the information will react to the explanation by asking about places she does not know, the giver of the explanation might ask for confirmations of understanding, etc. The structure of the interaction will be associated with the structure of the route, in a manner common to task-oriented dialogues. When studying interaction of this kind, the interplay of communicative actions (questions, answers, feedback) with the type of information exchanged (landmark descriptions, motion instructions, etc.) is central and gives us a way of studying how information is grounded between participants.

12.2 A Corpus of Explanation Dialogues and its Annotation

To collect empirical data about the communication of spatial knowledge we designed the following experiment: we recorded conversations between two speakers on both ends of a telephone line (to restrict the experiment to verbal communication). We had two sets of subjects. Each one in the first set was supposed to give an explanation of how to get to his/her location (an apartment in the city centre) to a subject in the other group, who was located at another apartment not too far from the first one, so that it could be reached on foot.

A 'giver' was always associated with an unfamiliar 'receiver' to avoid the use of personal mutual knowledge. To ensure that all conversations were realistic and all subjects motivated by the task, everybody in the second group was invited for a drink at the first apartment a few days after the experiment was conducted. A set of

21 dialogues was collected, comprising a total of 747 speech turns that were divided into 1235 segments/dialogue acts. All the examples in the chapter are extracted from this corpus and presented with their location in the corpus (Prévot, 2004).

The segmentation of the speech turns was made in two passes. During the first pass, classical syntactic units such as sentences or independent propositions were isolated. However the peculiarities of oral syntax resulted in bad inter-annotator agreement on these boundaries. Moreover, the analysis appeared all the more difficult as many units with clearly distinct functions were clustered together. To address these issues, another round of segmentation was performed, mainly refining the existing units into simple clauses and homogeneous functional units (e.g. '*ouais ouais ouais*' will be treated as one unit).

Dialogues were then annotated based on the following principles: each communicative act was labelled with the form (e.g. assertions, questions) and the function of the act in the context. Moreover, a precise description of the type of task-related information that was involved was needed. If landmarks are central to the route structure, it is necessary to see at what levels they appear and in what types of dialogue acts. This required a more precise annotation scheme than general ones such as DAMSL or Switchboard-DAMSL (Jurafsky *et al.*, 1997). The scheme from the Maptask corpus (Carletta *et al.*, 1997) was not used either because it was in the area of dialogue acts relating to acknowledgements that the most confusion between Maptask annotators was found (Kowtko, 1996) and it is precisely acknowledgements that constitute the type of dialogue acts the present

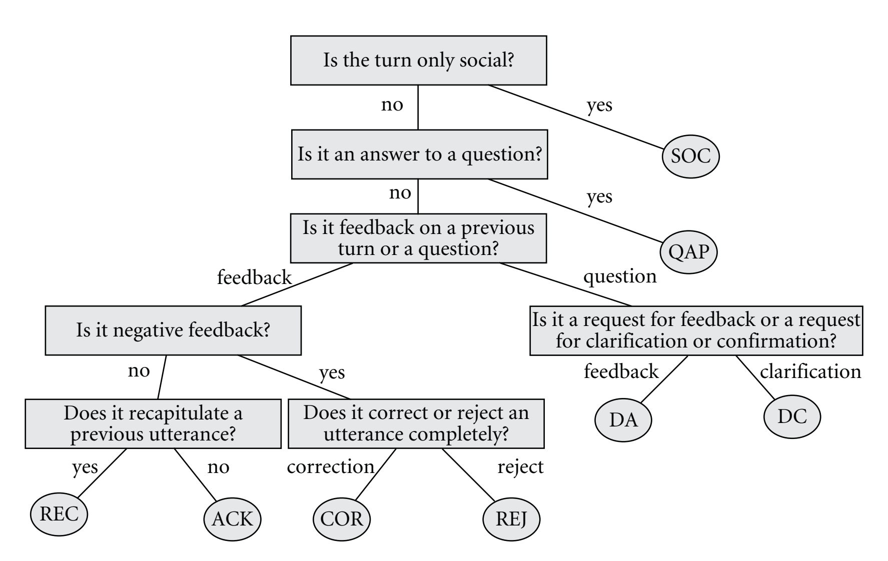


Fig. 12.1. Decision tree for the annotation of dialogue acts with a communicative function.

study focuses on. The surface part of the annotation was simply a distinction between assertions, questions (yes–no, wh-, or disjunctive questions), requests, and disfluencies. This annotation was carried out on the basis of the syntactic surface form and the intonation. The function-related part of the annotation scheme deals with task-based dialogue acts and communicative acts. Communicative acts were divided between answer, acknowledgement, recap, reject, correction, request for feedback, request for clarification, social act, and indeterminate. The indeterminate tag is not included in the decision tree but is actually present at each leaf. It is the default tag that is used when no other tag is applicable. Figure 12.1 shows the annotation decision tree, which is a subpart of the global tree when it is clear that the relevant turn has a communicative function.

12.3 **Dialogue Acts and Types of Spatial Descriptions**

12.3.1 The structure of route explanations

Denis and Briault (Denis, 1997; Denis and Briault, 1997), in their study of purely verbal monologic route descriptions, offer the following structure for explanations of a route:

1. Locate the receiver at the starting point. This means definition and anchoring of the starting point by the receiver and orienting of the receiver.
2. Start the progression ('do that').
3. Introduce a landmark. The landmark indicates the end of a step and helps orient the receiver again.
4. Orient the receiver.
5. Repeat 2, 3, 4 until the endpoint is reached.

They assume that landmarks are introduced in the order of the progression along the route. Moreover, they divide orientation dialogues into three phases: opening, orientation, and closing. Golding and colleagues (Golding *et al.*, 1996) claim that a question about orientation actually has two parts: how to proceed to a place and how to identify the place (if it is not known in advance). This last part is important, as the first one generally leads to a zone around the endpoint (in our corpus, it was a flat located in a small, not well-known street) whose granularity depends on the knowledge shared by speakers (Tomko and Winter, 2006). Another difference is that the starting point of the receiver might not be known by the giver, and there is also a need for explanation. In some cases, speakers use a strategy different from these: they just try to find a landmark near the endpoint that the receiver might know and build a simple route from there.

12.3.2 Contributions to route explanations

We will now have a more precise look at the basic elements of a description that constitute the semantic side of our dialogues. We take the classification of Denis (1997) as a basis (itself a refinement of Klein, 1982, and Riesbeck, 1980): prescriptions without landmark, prescriptions with landmarks, introductions of landmarks, descriptions of landmarks, and comments; and complete it with two types of dialogue acts: positioning and precision of a description. Percentages are given relative to the number of speech acts that are task-related (about half of the total number of 1235 speech acts); a slash (/) in examples indicates a speaker change.

- prescriptions without landmark (IOL): 5% of task-related speech acts in the corpus (17% in Denis). These can be either simple instructions (e.g. '*tu continues*' [*you go on*]) or reorientations (e.g. '*tu tournes à gauche*' [*you turn to the left*]).
- prescriptions with landmarks (IWL): 16% of task-related speech acts (33 % in Denis). These can make use of previously introduced landmarks, described with definite descriptions (proper names, demonstratives, etc.) or introducing new landmarks in the process. (e.g. '*tu traverses la rue de Metz*' [*you cross Metz Street*]). Finer semantic distinctions can be made by using a typology of actions or motion predicates. Denis has five types: go-to/go-towards X, take X (a street), go out of X, cross X, go past X, turn at X.
- introductions of landmarks (IL): 20% in proportion (36% in Denis). These are static descriptions of the form '*il y a X*' [*there is X*], '*tu verras*' [*you will see*]. For example: '*il y a un hôtel de police à cet endroit-là*' [*there is a police station at that place*]. Indefinite or definite descriptions can be used when the speaker assumes the landmark should be known by the other speaker.
- descriptions of landmarks (DR): 47%, the most important class (vs. 11% in Denis). The interactive nature of the corpus is the most straightforward explanation for this: speakers spend much time making sure the landmarks they used are identified and produce many utterances which return to previously introduced entities (e.g. '*c'est celle qui va des boulevards vers Rangueil*' [*it's the one that goes from the Boulevards to Rangueil*]).
- comments (COM): 6% of speech acts, mostly used to describe the task itself: '*tu verras c'est facile*' [*it will be easy*].
- positioning (LOC): 4%. The positioning label (LOC) is used mainly for relative clauses and adjuncts providing the localization or the situation for the next piece of information (usually a prescription or a landmark introduction). Such acts can also correspond to isolated phrases that could be part of a larger turn that would for example introduce a landmark or a prescription, but that are interrupted by another speaker: '*après le carrefour/oui/tu verras X*' [*after that junction /yes /you will see X*], but they can also be dislocations that are not finished '*quand tu es à X /je ne connais pas X/quand tu es à Y tu*

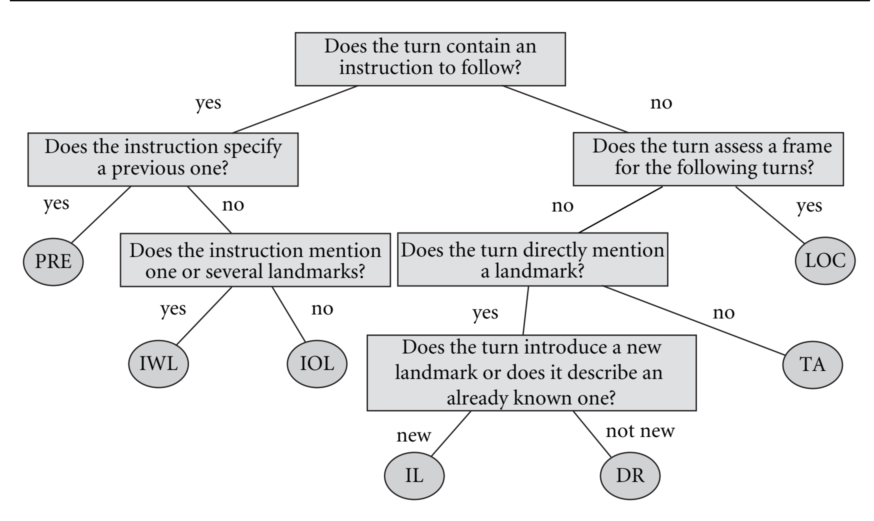


FIG. 12.2. Decision tree for task-related dialogue acts involving spatial descriptions.

tournes à gauche' [*when you are at X/I don't know X/when you are at Y you turn left*].

- precision of prescriptions (PRE, 2%). These give information about a preceding prescription; an example is '*tu remontes la rue X/mhmm/vers Y en fait*' [*you go up the street/mhmm/towards Y actually*].

Remaining task-related acts were labelled TA. These are constituted by prescriptions or descriptions related to the task but not to the route, such as the characteristics of the button for the doorbell, or an explanation of where the flat is once you entered the building.

12.3.3 A decision tree for route explanation contributions

The most relevant factor for the study, from a semantic point of view, is the presence or absence of a spatial entity, a landmark, because grounding of information depends almost exclusively on agreeing on these landmarks. People know or do not know some of the referents introduced, but usually have no problem following or planning an action once this is clear. So it should be expected that there are different ways of grounding classes that may involve new referents (introduction of landmarks, prescriptions with landmarks, most descriptions of landmarks), acts that don't involve new referents (prescriptions without landmarks, positioning, precision) but are part of the route, and acts that just describe the task itself (comments).

Figure 12.2 shows the decision tree for dialogue acts that are task-related and involve spatial descriptions. The kappa measure, advocated by Carletta (1996), was used to estimate inter-annotator agreement (the measure is balanced for chance

agreement). A good agreement is usually considered at 0.8, while a kappa between 0.6 and 0.8 is considered acceptable. Here, the kappa measure was approximately 0.8. There was the most confusion between the annotators for IWL and IL, because it was not clear in some cases if a turn was an instruction or not. Another source of discrepancies was the unclear status of many turns as polar questions or as mere assertions (because the syntax is often the same in modern oral French), leading to different labels for the following turns (either answers or acknowledgements). This prompted a study of this distinction, also taking intonation into account (Safarova *et al.*, 2005).

12.4 **Different Types of Acknowledgement**

12.4.1 Different functions for positive feedback

Feedback effects range from full rejection to full acceptance. Here, the focus is on positive feedback (or acknowledgement). Acknowledgements are sometimes mixed with back-channels. Back-channels are speech units that are uttered 'by the listener' or, to be more precise, uttered by the participant who does not have the floor. Acknowledgements are often back-channelled, but in theory any kind of contribution may be. In consequence, only the positive feedback aspect was considered, whether back-channelled or not. The speaker uttering the response might have heard, understood, or agreed on the target of the acknowledgement. It is not often clear what factors take part in defining this level of acceptance, nor how they interact with the other functions of feedback. Therefore, during the annotation task, there was only one kind of acknowledgement, but it was necessary to be able to distinguish different levels a posteriori. Feedback is associated with the establishment of different kinds of objects: propositions and/or their truth on the one hand and referents and their identity on the other hand. This can be further refined into four phenomena:

- grounding (understanding, settling a proposition) is not related to the truth of utterances or the acceptance of an order. It is just a coordination between the speakers on what has been said, not necessarily its truth nor acceptance of the information, as exemplified in (1). In these examples, G stands for the Giver of the explanation, R for the receiver.

  (1) G_{21}. et au coin y'a un restau de tapas.
  and at the corner there is a tapas restaurant.
  $R_{22.1}$. ouais. *yeah.*
  $R_{22.2}$. je vois pas du tout. *I don't see it at all.*

- accepting (agreement), as opposed to grounding, leads to the acceptance of the truth of the information agreed on or at least it leads to the acceptance of the information with respect to the current purpose. See example (2).

(2) G_{21}. tu vois la première rue à droite avant d'arriver sur la place en fait.
you see the first street on the right before arriving at the square in fact
R_{22}. ouais d'accord. *yeah ok.*

- anchoring (establishing a referent) is finding an internal anchor in one's beliefs for what has been said by the other speaker (e.g. a common referent has been found). See example (3).

(3) $G_{41.3}$. c'est au 27 rue des Polinaires.
it is at 27 Polinaires Street.
$R_{42.1}$. ouais je vois. *yeah I see.*
$R_{42.2}$. en fait c'est rue des Polinaires.
in fact it is Polinaires Street
$R_{42.3}$. d'accord. *Ok.*
F_{43}. voilà. *that's it.*

- closing is terminating a span of discourse as a subdialogue successfully or not (See F_{43} in 3). It was decided to include it in the list because (i) such closings after successful exchange or not are still agreed upon among the speakers and (ii) they concern mainly positive closures.

However, systematically finding the function of a given feedback utterance is not an easy task. There is no direct entailment between them except maybe that acceptance requires grounding. Accepting an utterance also requires, most of the time, to have anchored it beforehand. Anchoring a sentence means establishing all the referents it contains.

12.4.2 The targets of positive feedback

Another factor that was looked at is the function of the previous utterance in the context (its relational function). The data presented in Table 12.1 distinguish coarsely between task-related assertions describing an itinerary: introduction of landmarks or description of landmarks, instructions and communication management turns that are not related to the task (mainly feedback). These

TABLE 12.1. *Targets of acknowledgement markers by general dialogue act types*

Marker	ouais	oui	ok	d'accord	voilà	mhmm	bon	je vois	others	total
Landmark	83	26	30	64	14	9	9	17	4	256
Instruction	38	10	3	18	2	14	0	2	0	87
Feedback	23	6	18	18	34	6	19	1	3	128
Task	21	12	7	23	2	9	3	2	0	79
Total	165	54	58	123	52	38	31	22	7	550

distinctions are important with respect to the difference between acceptance and anchoring, since anchoring is mainly about landmark management. The utterance concerned by the management of landmarks mainly aims to anchor the referents they include, while instructions need to be accepted. Anchoring underlies the establishment of 'managing referent utterances' and grounding the establishment of 'instruction utterances'.

12.4.3 Lexical markers of positive feedback

A set of discourse markers (DM) associated with positive feedback in French have been isolated. Discourse markers have been studied for some time (Schiffrin, 1987) but there are only a few studies concerned specifically with lexical feedback markers; see, however, the study of back-channels '*mhmm*' in Gardner (2001), the comparison of '*mhmm*' and '*ok*' in Bangerter and Clark (2003) and the recent volume on lexical markers by Fetzer and Fischer (2007). Studies on French (such as Roulet *et al.*, 1985; Rossari *et al.*, 2004) are not focused on feedback issues. We add to the DM analysis the observation of informationally redundant utterances that help participants infer acceptance, as shown in Walker (1992). In order to evaluate the functions of the different markers, the correlation between each marker and the different roles that seemed relevant was statistically tested. Only the most interesting with respect to spatial information are shown here, namely the nature of the target of the feedback utterance. A different test was made for each marker with respect to the following possible types of targets: instructions (IWL, IOL), landmark introduction or description (IL, DR), other task related acts, and feedback turns. In each case a Fisher exact probability test was used on the contingency table (marker is present in feedback, marker is not present in feedback) against types (landmark, instruction, other task, feedback). The four broad categories used result from a simple grouping of the categories introduced in Section 12.3.2.

Table 12.2. *Correlation marker vs. target of feedback*

Marker	p	landmark	instruction	task	feedback
ouais	0.0004	6.2	11.9	−2.7	−15.4
oui	0.0651				
ok	0.0550				
d'accord	0.0346	6.75	−1.46	5.33	−10.63
voilà	0.0000	−10.2	−6.23	−5.47	21.9
mhmm	0.0004	−8.69	7.99	3.54	−2.84
bon	0.0000	−5.43	−4.9	−1.45	11.79
je vois	0.0276	6.76	−1.48	−1.16	−4.12
others	0.3976				

Table 12.3. *Breakdown of marker use by speaker*

Count	French marker	by giver	by receiver	English equivalent
141	oui, ouais	34	107	yes/yeah
67	d'accord	18	49	ok, I see
47	voilà	38	9	exactly, that's it
37	ok	9	28	ok
29	mhmm	8	21	mmmh
18	bon	10	8	now, ok, well
14	je vois	1	13	I see
12	repeat	8	4	—
22	other	—	—	—

Table 12.2 indicates for each case the level of significance (p value) and, when this is under 0.05, the difference with respect to an expected random distribution (marginals) in order to hint at what type was more likely to have influenced the test. In other words, a positive difference indicates a correlation between a marker and the role of the target, and the higher the difference, the further the presence of the marker is from a random distribution.

Our hypothesis is that each marker fulfils a different function within the acknowledgement process, and is thus likely to be associated differently with various dialogue acts. This assumes that these acts are involved differently in the process of grounding, accepting, or anchoring information (although it is difficult to establish how in every particular case without access to the speakers' mental states). From these results, it seems the hypothesis was valid: most markers provide feedback more often on certain classes of semantic acts. The markers can be grouped according to these acts: involving referents (*ouais, d'accord, je vois*), instructions (*ouais, mmhm*), or more typically used on feedback (*voilà, bon*). It remains to be seen to what extent these groups correspond exactly to different levels of acknowledgement. The specific targets of acknowledgements including '*voilà*' or '*bon*' suggest these DMs to be very efficient cues for detecting closure. The '*mhmm*' and '*ouais*' targeting instructions are very weak acknowledgements that are basically elicitations (often back-channelled) for the next instructions. Finally, '*je vois*' targets primarily landmark-related utterances confirming their landmark-anchoring role. These different functions explain the uneven repartition of DM use among roles (giver/receiver) as exhibited in Table 12.3.

12.4.4 Correlation with speaker roles

The asymmetry between the two participants allowed for the investigation of which markers were correlated with speaker roles. These roles can be associated

with initiative or with competence concerning the issue currently discussed. Unsurprisingly, the general pattern of the receiver acknowledging the information provided by the giver is confirmed by the data. Some phenomena nevertheless provide some finer conclusions. The DMs '*voilà*' and '*bon*' confirm their closure marker roles. Contrary to all other DMs, they are mostly produced by the giver who has authority over the information. They are used for acknowledging receiver feedback in order to close a given topic.

12.5 Conclusion

The goal of the present study was to determine how different types of spatial information (landmark management, motion instruction) are related to different kinds of positive verbal feedback. Different kinds of feedback were identified that are supported by our data and, looking at various lexical cues in utterances, the type of dialogue act targeted by the feedback. Refining dialogue acts with respect to the specificities of direction-giving shows the importance of the distinction between managing referents in the dialogue history, giving instructions, and managing coordination in the interaction. It provides a good empirical basis for the characterization of different feedback acts (grounding, accepting, anchoring). Finally, it gives some validity to the difference in cognitive status of the various types of spatial expressions used by speakers in route-explanation dialogues.

12.6 Acknowledgements

The work presented in this chapter has been partially completed while Laurent Prévot was supported by a research grant from Trentino province and then by a research grant from the Institute of Linguistics at the Academia Sinica.

13

Telling Rolland Where to Go: HRI Dialogues on Route Navigation

SHI HUI and THORA TENBRINK

13.1 Introduction

Within the DFG Collaborative Research Center SFB/TR 8 'Spatial Cognition', one prominent research aim is to achieve effective and natural communication between humans and robots about spatial issues. While some of the work so far has addressed spatial instruction scenarios in which a robot was instructed to move towards one of several similar objects present in a scenario (e.g. Moratz and Tenbrink, 2006), the present aim is to enable previously uninformed users to 'tell Rolland where to go' via a suitable dialogue model. 'Rolland' is the name of the Bremen autonomous wheelchair (e.g. Lankenau and Röfer, 2000; see Figure 13.1 below), a mobile robot equipped with sensorial and conceptual functionalities, which we use to investigate how humans communicate linguistically about route navigation, particularly focusing on dialogue processes between humans and robots.

With respect to this target scenario, considerable progress is currently being made in related fields in several crucial respects. Technologically, major advances are being achieved in human–robot interaction by combining language processing with multimodality (e.g. Skubic *et al.*, 2004) and distributed modular architectures (e.g. Spexard *et al.*, 2006), sometimes integrating ontological knowledge with the robot's sensor-based world representation (e.g. Kruijff *et al.*, 2007). These robotic systems are highly advanced and work well with users familiar with the system's functionalities. Also, major advances have been documented within dialogue systems research (e.g. Lemon *et al.*, 2003; Wahlster, 2006). Finally, some research groups focus on the automatic treatment of route instructions given to robots (e.g. Kyriacou *et al.*, 2005; Ligozat, 2000; Tschander *et al.*, 2003). Such approaches are well-founded in terms of cognitive appropriateness in that they rely directly on humans' natural conceptions and linguistic representations in route navigation. Notably, there are major differences concerning the requirements for online (accompanying) route directions, which may rely entirely on simple directionals such as 'go left', and in-advance route directions that involve complex spatial



FIG. 13.1. Rolland—The Bremen autonomous wheelchair.

descriptions derived from mental representations, which need to be mapped to the real world (Habel, 2003).

The specific features of route communication in a dialogic situation even between humans (rather than between humans and robots) have seldom been investigated; however, a thorough and fine-grained analysis of a Maptask dialogue is presented by Filipi and Wales (2004), who show that shifts of perspective (i.e. the conceptualization and linguistic representation of the spatial situation) occur systematically as required by the demands of the interaction. Their recent study (Filipi and Wales, this volume) highlights how the verbs *come* and *go* relate to such perspective shifts, associating the route-internal perspective (referring to the movement through space) with the verb *go*, and the route-external perspective (referring to the map rather than the world it represents) with *come*. Such regular variability in speakers' representations needs to be accounted for

in human–robot interaction, especially since there may be additional reasons for humans to produce unexpected utterances when interacting particularly with robots (cf. Holzapfel and Gieselmann, 2004). An advanced dialogue system should be capable of accounting for systematic communication problems such as these dynamically when designed specifically for the requirements within a certain application field. This is the basic idea that we address in this chapter, drawing on experiences with our own dialogic robotic system described in Ross, Shi *et al.* (2005).

The precise experimental scenario we have in mind is as follows. Users seated in the wheelchair are asked to move around in the university building, explaining the most important rooms and places (landmarks) to Rolland as they go. They are told that the robot completes its internal map in this way. After this, the users are asked to instruct Rolland to navigate to a specific place where they have previously been together with the wheelchair. Rolland is then supposed to move autonomously to the intended goal. At this point, a number of communication problems may occur because of mismatches between the robot's knowledge and the user's linguistic representation. Thus, on the one hand we are interested in users' spontaneous strategies in instructing the robot for the navigation task; on the other hand we investigate how clarification dialogues can be initiated by the robot in an effective way in cases of communication failure.

Our general approach is as follows. Our robot is equipped with a conceptual route graph, based on findings in the literature concerning how humans represent spatial settings (Section 13.2). In order to establish the relationship between the implemented map and users' spontaneous references to a particular spatial environment, we collected empirical data in a scenario in which the robot wheelchair did not act autonomously (Section 13.3). We used the information provided by the participants in order to augment the robot's internal map, and identified a number of conceptually problematic aspects concerning the relationship between the robot's knowledge and the humans' linguistic choices (Section 13.4). We propose to approach these problems by employing an adequate dialogue model that is capable of handling conceptual mismatches similar to the alignment and repair strategies in a natural dialogue (Section 13.5). The next logical step is the empirical testing of the implemented dialogue system with uninformed users, leading to an improvement of the model.

13.2 Cognitive Map: a Representation of Rolland's Spatial Knowledge

Dialogue systems have typically been used previously in information query systems, in planning systems, or in intelligent robotic systems. Since almost all dialogue systems use specific domain knowledge to interact with users, the

representation of such knowledge determines not only the content but also the way the system interacts with those users. Thus, the naturalness and the efficiency of a dialogue system depend largely on the representation of domain knowledge.

Humans' cognitive models of spatial situations do not directly mirror the actual relations measurable in the world. This can be seen in the ways humans represent their spatial concepts externally—for example, in verbal route directions or sketch maps (Tversky and Lee, 1998). Such externalizations highlight how mental representations are schematized, systematically simplified, and often distorted (Tversky, 2003; Talmy, 2000); also, they reflect a range of different conceptualizations or 'perspectives' (Tversky, 1996; Filipi and Wales, 2004). A number of investigations (e.g. Denis, 1997; Michon and Denis, 2001; Klippel *et al.*, in press) classify the essential ingredients of a spatial description: they contain information about orientation and direction, as well as decision points, landmarks, and continuations of movements. Our own empirical data (see next section) basically conform to this classification.

Within Artificial Intelligence, such mental conceptualizations of space are formalized and modelled in the subfield of qualitative spatial representation and reasoning. This kind of automation can make valuable predictions about human spatial behaviour even in cases where a precise quantitative representation is not available or is too complicated to be treated computationally (Cohn *et al.*, 1997). Spatial knowledge can be represented in a number of ways (see, for instance, the conceptual layer proposed in Kruijff *et al.*, 2007); in our approach we combine Freksa's Qualitative Spatial Calculus using orientation information (Freksa, 1992) and the Route Graph (Werner *et al.*, 2000) to represent Rolland's conceptual spatial knowledge for communication with users.

Route Graph. The general concept of Route Graphs is suitable for modelling navigation of various kinds of agents in diverse scenarios (Krieg-Brückner *et al.*, 2005). They can be used as metrical maps combining sensory input with metric computational models for controlling the navigation of robotic agents in the environment; likewise, they can also be used at the cognitive level to abstractly model humans' topological knowledge while they act in space. Importantly, Route Graphs integrate different routes as well as different kinds of information into a coherent graph-like structure.

The starting point for the definition of Route Graphs is basic graph theory, yielding nodes and edges. A *node* of a Route Graph, called a *place*, has a particular position and specification ('reference system') which may be related to a global reference system. An *edge* of a Route Graph is directed from a source node to a target node and is called a *route segment*. It has three additional attributes: an *entry*, a *course*, and an *exit*. The information associated with these attributes is specifically defined in each Route Graph instantiation. For example, in a Route Graph at the metrical level, an entry or an exit can be an angle defined by its degree value with respect to a global 2-D geometric system; the course is simply characterized by metrical data about length and width. In contrast to this, an

entry or exit at the cognitive level may contain qualitative orientation information (such as *to the left/right*), while the course consists of the path between two reorientations. Finally, a *route* is defined by the conjunction of route segments from one place to another. More details of the general concept of Route Graphs are given in Werner *et al.* (2000) and Krieg-Brückner *et al.* (2005).

Qualitative spatial calculus using orientation information. Freksa's calculus for qualitative spatial reasoning (Freksa 1992) uses orientation information given by two points in 2-dimensional space: the start point and the endpoint of a movement. By combining the *front/back* and the *left/right* dichotomy, eight meaningful disjoint orientation relations can be distinguished (e.g., *straight-front*, *right-front*, *right-neutral*, etc.). Freksa (1992) gives the following simple example to show how a location can be determined on the basis of our own location and other locations we know, which is an important application of orientation-based qualitative spatial reasoning. Imagine walking from location a to location c and reaching the intermediate location b. We can describe orientation and distance of location c qualitatively with respect to the segment between location a and b as an oriented line, denoted as the *vector* **ab**, that is, we compare the vectors **bc** and **ab** with respect to their orientation. At position c, we can similarly compare a further vector **cd** with the previous one, **bc**. By means of an inference step, we can then determine the goal location d in relation to the initial vector **ab**.

Combining these two approaches we gain what we call the *conceptual route graph* (Krieg-Brückner and Shi, 2006). *Conceptual* here reflects the fact that the model is designed to partially represent human navigation knowledge, thus enabling humans to interact with robots more naturally. This stands in contrast to the more typical metric maps often used by robots to carry out navigation tasks. The conceptual route graph consists of the four basic types *orientation*, *route mark* (the Route Graph equivalent of a landmark), *place*, and *vector*. Moreover, a set of relations between these types is defined. For example, *at* defines the relation between place and route mark, *on* the relation between place and vector, and *ori* the relation between two vectors or between a vector and a place. In a conceptual route graph, entries and exits of route segments are given as orientation relations, while courses of route segments are vectors. Figure 13.2 gives a conceptual route graph of the floor where the empirical study described in the next section took place; the background shows the floor plan. There are several reasons for using the conceptual route graph. First, Freksa's qualitative spatial calculus is well defined and provides an efficient way of representing and reasoning about causal actions and topological relations, while the Route Graph provides concepts like route segments, routes, and a number of spatial relations. Second, our empirical studies show that the combination is powerful enough to cover the larger part of the users' descriptions used in the communication with the robot during a spatial navigation task (see next section), e.g. *at(staircase, p) and ori*($\mathbf{p}_0\mathbf{p}_1$,*p, left)* represent 'the staircase is to our left' if the last movement is $\mathbf{p}_0\mathbf{p}_1$. Third, in practice this representation supports an efficient mapping to the robot's spatial representation

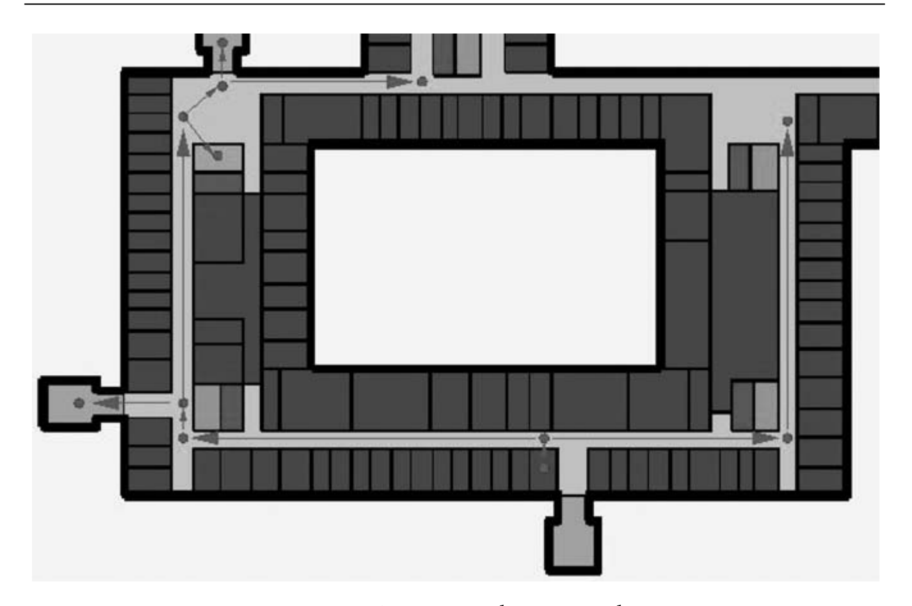


Fig. 13.2. A conceptual route graph.

at the metric level for real navigation, and vice versa (Shi, Mandel, and Ross, 2007). In the next section we describe our empirical study in which we collected human users' spontaneous ways of referring to the environment implemented in the conceptual map.

13.3 Spatial Reference in Route Navigation: Experimental Study

Method. A decisive design feature of our methodology (cf. Fischer, 2003) is to keep users basically uninformed about the robot's capabilities, and to suggest a realistic (though partly illusory) human–robot interaction scenario to them that is suitable for triggering the kind of utterances that the system should understand in the future. This method ensures that speakers' spontaneous linguistic strategies and choices can be investigated, in contrast to approaches in which existing human–robot interaction systems are evaluated by people who are entirely familiar with the robot's capabilities. In particular, we wish to account for the fact that speakers react differently depending on the addressee they are confronted with, as shown, for instance, by Schober (this volume). Since robots differ fundamentally from humans, the humans' linguistic behaviour in human–robot dialogue cannot be predicted from human–human dialogue research.

The 23 German and six English native speakers participating in our study were told that the robot wheelchair was capable of processing information given

verbally in their native language, and that it would then use it for later navigation tasks. However, in this study, Rolland never gave any kind of response in reaction to the users' utterances, or feedback that it had actually understood the message. This is the case because in this exploratory phase Rolland's actual functionalities were not used at all, contrary to what we told the users. The users' utterances were recorded, transcribed, and analysed qualitatively with respect to their relation to the robot's functionalities, in particular with respect to the range of linguistic user strategies and conceptual patterns. No assessment of relative frequencies or other statistical measures were undertaken because of the study's open design in which users were free to solve the task entirely according to their own preferences. Here we focus on those results that are directly relevant for the development of dialogue systems.

Procedure. The experiment consisted of four phases, the first two of which took place within a room and are not analysed further here (but see Tenbrink, 2006). In the third phase, the users, who were seated in the wheelchair, were asked to steer Rolland around the hallways in the university (see Figure 13.2 above) along a prescribed route and talk to it, explaining the spatial relationships that Rolland would need to know in later tasks. In the fourth phase of the experiment, the users were asked to instruct Rolland to move to a particular place called the 'Stugaraum' or 'Stugeneck' (a students' common room). We told them that Rolland would now be able to navigate autonomously to this place, which of course was not the case since the robot could not integrate the information directly.

Results. In the third phase, speakers displayed different individual priorities and strategies, and they produced descriptions at various levels of granularity. For example, they named entities along the route and described their spatial location, such as 'the first door to the right, here is the copy machine'. In other cases, they described their current direction of movement, for instance (numbers indicate pauses in seconds): 'left (5) straight ahead (20) and slightly left (3) slightly right (17) hard left', and the like. In the fourth phase, users produced in-advance route directions towards the goal, again using various strategies and employing different levels of granularity, but generally attending to the kinds of information they had previously given the robot. An example is, 'we gonna turn around, go out the door and make a left, continue down the hallway straight through one door, we're gonna make one more left, and continue down that hallway to the stugaroom'. For present purposes we do not attempt a detailed linguistic analysis of these utterances, but we aim to establish their relationship to the implemented map and Rolland's potential ability to deal appropriately with users' spontaneous choices of route instructions.

Relation to Rolland. The results of the third experimental phase were used after the experiment to augment the robot's internal map, that is, they were added into the conceptual route graph. For example, Rolland 'learned' the location of the mailbox room by the description 'hier rechts von mir ist die Tür zu dem Raum

mit den Postfächern' (here to the right of me is the door to the room with the mailboxes), represented as *{ori(v, p, right-neutral), at(theMailbox, p)}*, where v is the vector including the current place and orientation, p the place to the right of v, and the route mark 'the mailbox' is at p.

Our main interest at present concerns the in-advance descriptions collected in phase four. The linguistic data collected in this phase were used as a data pool to identify potentially problematic conceptual phenomena (cf. Section 13.4). Some of the utterances could not be understood or interpreted by Rolland according to its current knowledge representation, even after augmenting the internal representation. In the following we describe a range of generalized problem areas derived from our data. These need to be addressed for human–robot interaction in our scenario to succeed. For this purpose, we present the first version of our dialogue model in Section 13.5.

13.4 Representation Disparities

Reference resolution. Previous research has shown that speakers' assumptions concerning what robots can or cannot understand crucially determine their linguistic strategies towards the robot (Fischer and Moratz, 2001). In our scenario, users could assume that the robot had been informed about the position of entities during the prior exploration; however, it is not the case that they reliably used the kind of vocabulary that they had used in the exploration phase, although they quite obviously tried to do so (in one case, a speaker even tried to imitate her own previous pronunciation: 'Rolland I want you to go to the stugaroom, or did I say sh-tugaroom'). Nevertheless, there were some occurrences of new vocabulary in the fourth phase. Moreover, the user may refer to entities or actions in the real world that are present at some place in the robot's internal representation but not at the place the user refers to, which leads to a representation mismatch. This might be due either to the users' distorted representation in memory, or to a mismatch between the robot's internal map and the current state of the real world.

Conceptual knowledge. As stated in Section 13.2, Rolland's spatial knowledge is represented as a conceptual route graph, which is suitable for the representation of most humans' spatial descriptions in an office-building environment. However, a few utterances also contained references to regions rather than route-like formats, such as 'vorbei an den Haupttreppen—also gar nicht in den Bereich reinfahr'n' (past the main staircase—that is, do not go into this region). An examination of a different corpus, the IBL—Instruction Based Learning corpus (Kyriacou *et al.*, 2005)—indicates that, in other scenarios, region-based references may be more common, as shown by utterances such as 'you are in the parking space'. Thus, such representations may need to be integrated into the conceptual map. Such representation mismatches of formats and features could be due to a principled

incompatibility of the representations. This would be the case if communication failed repeatedly because of such mismatches in a real human–robot interaction situation. For present purposes, we focus on the ways in which such representation problems—regardless of their source—can be handled efficiently via dialogue.

Underspecification. Many utterances contained vague and underspecified representations, such as 'n kleines Stück geradeaus' (a bit straight on) instead of specifying an exact distance that the robot should cover, or a destination towards which it should move. Some users simply repeated direction instructions, e.g. 'geradeaus geradeaus geradeaus (…) irgendwann nach links' (straight on straight on straight on (…) eventually to the left). Sometimes users employed expressions like 'ungefähr' (roughly), 'weiß nicht' (don't know), 'irgendwo' (somewhere), reflecting uncertainty. Such utterances cannot directly be represented in Rolland's conceptual route graph. Our solution to this problem is the introduction of *place variables* and *orientation variables*. For example, 'irgendwo links oder rechts kommt dann dieser Turm' (somewhere on the left or on the right comes this tower) is then represented as *ori(x, p, right-neutral)* or *ori(x, p, left-neutral)*, where 'this tower' is at place p, and the place variable x stands for some place on the path such that p is right or left of it. Such variables can often be resolved according to existing relations in the conceptual route graph and further route descriptions. In the case where a variable cannot be resolved, a subdialogue for requesting more information should be generated (Section 13.5).

Granularity. Instructions occurred on different levels of granularity, either referring directly to the goal, or listing route marks, or specifying each turn, etc. If the conceptual route graph contains the information on the level of granularity employed by the user, then the instruction maps to the robot's knowledge. Problems arise in the two other kinds of possibilities. The utterance may contain only a coarse description to a goal that the robot does not know how to reach, or it may contain only low-level information that would lead to a continuous need for interaction, in spite of the fact that the robot has access to higher-level information in its knowledge representation. Thus, when a user only employs directional information such as 'rechts' (right) or 'geradeaus' (straight on), corresponding route segments cannot be created. Then, the communication situation would profit from a dialogue that induces users to employ the highest level of granularity that the robot is capable of dealing with.

Boundaries of tasks and actions. As stated in Section 13.2 route segments directed from source place to target place are basic components required to build routes in the conceptual route graph. Thus, identifying boundaries of tasks and actions is important for completing route segments. However, many utterances do not contain explicit specifications of spatial and temporal boundaries for tasks and actions, that is, when or where they start, and when or where they are completed. In some cases such boundaries were indicated in some way, either spatially or temporally, although such information may not always be useful to the robot if no

corresponding information is represented in the conceptual map. In other cases the boundaries need to be inferred, for example by investigating neighbouring instructions. A spatial end boundary is indicated whenever people specify a goal or subgoal, as in 'weiter geradeaus bis zum letzten Zimmer auf der linken Seite' (further straight on until the last room on the left side), or the length of a path is specified: 'dreißig Meter vielleicht' (maybe 30 metres). Information about time is usually vague, as in 'ziemlich lange geradeaus' (straight on for a fairly long time). In other cases an action ends when the next action starts, as in the spatial example 'immer geradeaus, das nächste Mal wenn's links geht, links' (keep straight on, turn left the next time when it's possible to turn left); and the temporal one: 'immer geradeaus, irgendwann links' (keep straight on, eventually left).

Temporal order. If the temporal sequence that the speakers have in mind for the completion of the task corresponds to the order of representation, then it can be directly represented as a route in the conceptual route graph because a route is defined as a temporally-ordered sequence of route segments. However, sometimes people go back mentally to add further information: 'dann auf dem Flur nach links fahren bis zur ersten möglichen Abzweigung nach links, also da kommt dann noch dieser Turm aber der interessiert uns nicht wir wollen also nach links' (then drive on the corridor to the left until the first possible turn off to the left, I mean, there is also this tower but this does not interest us we want to go left). In this specific case the temporal digression is also a digression with respect to the task, that is, the utterance contains further information about the scenario that is not directly related to the instruction (as indicated by 'this does not interest us'). In this experiment no utterances contained *instructions* in a distorted temporal order, which may mean that it can be taken as the default case that instructions are given in the intended temporal order. Using the conceptual route graph we can treat digressions as additional spatial relations on the route representation given in temporal order, but route instructions in a distorted temporal order could not be represented properly.

13.5 **Dialogue Modelling**

According to the above analysis, dialogic interaction between the user and the robot is indispensable in order to achieve a definitive route. We must therefore find appropriate ways of modelling such dialogic interaction. Here there are various approaches potentially to choose from. Grosz and Sidner's theory of discourse structure, for example, is composed of three components that collectively structure and analyse dialogues: the structure of utterances, the structure of purposes, and the structure of attentions (Grosz and Sidner, 1986). A further approach is Carletta *et al.*'s Dialogue Structure Coding Schema, which consists of three levels: dialogue move coding, dialogue game coding, and transaction coding (Carletta

et al., 1997). The major goal of these approaches, however, is to provide a theoretical foundation for the representation and processing of dialogues. In contrast, our focus in this work is on the formalization of dialogue control according to discourse patterns gained from empirical studies in order to enable flexible dialogue between humans and the robotic wheelchair. The method followed here is as a consequence based on the COnversational Roles (COR) model (Sitter and Stein, 1992). This is a generic situation-independent dialogue model which can be restricted or extended to cover precisely particular empirically motivated discourse patterns that are revealed during experimentation (cf. Tenbrink and Shi, 2007; Ross, Bateman, and Shi, 2005). Similar to *Information State* based approaches (e.g. Larsson and Traum, 2000) the model accounts for particular abstract states of information that may occur in ongoing dialogue. We focus on the following concrete consequences that may come about due to one or more of the issues listed in the previous section:

1. The robot does not know some of the concepts contained in the user's instruction.
2. The robot succeeds in identifying a direction of movement but not a destination. Thus, it does not know about the end boundary of the intended action.
3. The robot does not know how to interpret the user's instruction because of some kind of underspecification or some unknown spatial relations.
4. The robot recognizes that a destination has been mentioned but does not have access to information about how to get there, e.g. in the case of a granularity mismatch.
5. The robot detects an instruction in a distorted temporal order.
6. The robot finds that direction and destination, according to the instruction, do not match one another.

In the first two cases, the robot should request the user to give more information. To treat problems (3) and (4), the robot can make some suggestion according to its knowledge, such that the user can give more precise information or use a more suitable level of granularity, or it may request more information. If a conflict between the user's description and the robot's knowledge (problem 6) or a temporally distorted instruction is detected (5), the robot should inform the user so that the user can correct the description if there is indeed a mistake, or choose a new way of giving the instruction otherwise.

Figure 13.3 shows a recursive transition network representing a simplified dialogue model covering some of these cases. Here the interaction possibilities have been split into three subnetworks, the first describing the dialogic moves parameterized by participants, the second giving the possibilities for the user to give instructions and additions, or for the robot to ask for extensive feedback, and the

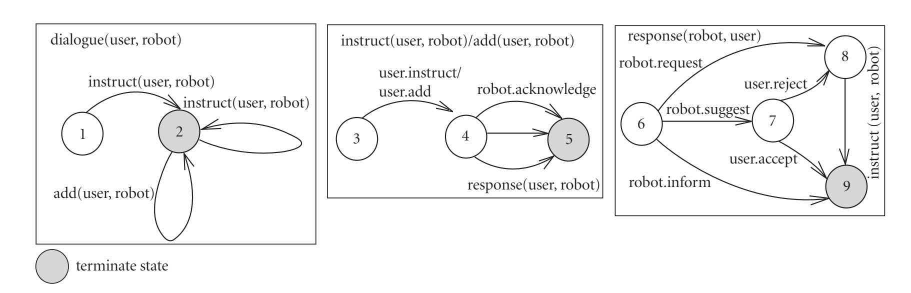


FIG. 13.3. The dialogue.

third focusing on the subdialogues generated by the robot to clarify problematic issues. For modelling dialogues using transition networks we define the intentions of dialogue participants as a set of dialogue acts. Here we use *instruct* for giving instructions; *add* for giving additional information; *accept* and *reject* for agreement and disagreement; and *request*, *suggest*, and *inform* for the robot to initiate subdialogues to clarify knowledge disparities and mismatches. In the transition network, parameterized moves, such as *instruct(user, robot)*, *add(user, robot)*, *dialogue(user, robot)*, and *response(robot, user)* are further defined as subdialogues. Moves like *user.instruct*, *user.add*, *robot.request*, and *robot.acknowledge* are basic dialogue moves—that is, moves that are directly implemented by other components as speech acts and actions.

The most important advantage of using Recursive Transition Networks to model dialogue in this way is the possibility of using existing computational methods, mechanisms, and tools to analyse such dialogue models. Furthermore, this approach allows for extracting dialogue models from existing dialogue systems, constructing dialogue models from empirical data, and enabling the analysis of model features, complexity, and coverage. When considering the construction of dialogue models from empirical data two questions often occur. The first question concerns the proper 'improvement' relation, that is, does an improved model contain the original one, to ensure that all discourse patterns discovered in previous stages are still covered by the new model? The second relates an instance dialogue to a dialogue model, that is, is a given dialogue an instance of a given dialogue model? Following the main idea in Shi *et al.* (2005), when represented as a recursive transition network, a dialogue model can be straightforwardly formalized in the specification language CSP (Hoare, 1985; Roscoe, 1998). Moreover, if we concentrate on dialogue actions, concrete dialogues can be abstracted to CSP specifications as well. By this means, actually produced dialogues can be formally compared with the possibilities intended by an abstract dialogue model. Then, using a standard model checker such as FDR (Formal Systems, 2001), we can proceed to *automatically* validate and analyse specified dialogues and dialogue models drawing on their CSP specifications.

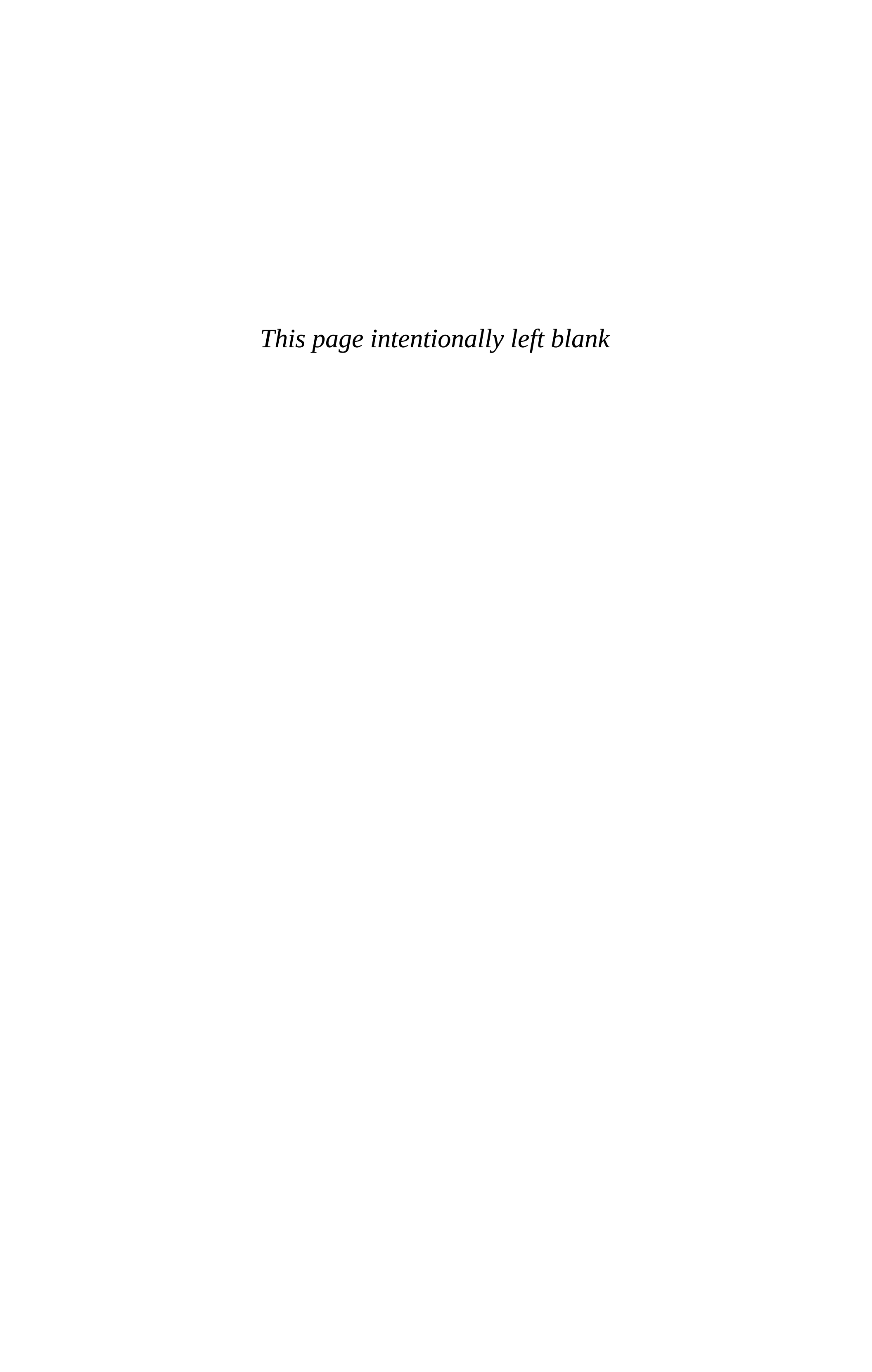
13.6 Conclusion and Outlook

Our approach towards natural and efficient human–robot dialogues for route-instruction tasks involves a cyclic approach iterating empirical investigation and dialogue modelling. Based on empirical results, we have presented a range of problematic aspects that can cause communication failure because of knowledge disparities between the robot's internal representation and the users' input, and suggested ways of addressing these problems via clarification dialogues. In Tenbrink and Shi (2007) we present a detailed analysis of a Wizard-of-Oz study specifically tailored to the intended functionalities of the robotic wheelchair Rolland, employing the first version of the dialogue model. Results show that our proposed model is successful in encouraging the user to provide missing information and to use a suitable level of granularity. However, new problems arise due to dialogic effects not covered by the first iteration reported here. Results such as these are used to improve the dialogue model in an iterative process. Furthermore, our dialogue model may be refined using empirical findings about how specific kinds of feedback influence dialogic conversation on spatial tasks. Muller and Prévot (this volume), for instance, propose a refined classification of positive feedback (e.g. grounding, accepting, anchoring).

The dialogue-modelling approach discussed in this chapter may be supported by a computational method for the analysis of dialogue models. In addition to the model construction process, the computational method has been applied to simulate dialogue models together with components for natural language processing, generation, and the application domain. We are now planning further empirical studies to evaluate the dialogue models developed in the current work with this simulator.

Acknowledgements

Most of all we thank Kerstin Fischer who took part in the preparation and realization of the studies discussed here. A number of other people within the projects of the SFB/TR 8 on Spatial Cognition have been involved in other aspects of the present endeavour: John Bateman, Scott Farrar, Udo Frese, Bernd Krieg-Brückner, Christian Mandel, Reinhard Moratz, Thomas Röfer, Robert Ross, Tilman Vierhuff, and others. Their participation in the development of the ideas and results presented here is acknowledged. We are furthermore grateful for the funding from the DFG for the projects I1-[OntoSpace] and I3-[SharC] of the SFB/TR 8 'Spatial Cognition', Bremen/Freiburg.


This page intentionally left blank

References

Abbeduto, L., Weissman, M., and Short-Meyerson, K. (1999). 'Parental scaffolding of the discourse of children and adolescents with intellectual disability: the case of referential expressions', *Journal of Intellectual Disability Research* 43: 540–57.

Allen, G. L. (2000). 'Principles and practices for communicating route knowledge', *Applied Cognitive Psychology* 14: 333–59.

Allen, J. and Core, M. (1997). 'Draft of DAMSL: Dialogue Markup in Several Layers', retrieved on 10 February 2008 from http://www.cs.rochester.edu/research/speech/damsl/RevisedManual/.

Allwood, J. (1995). 'An activity-based approach to pragmatics', *Gothenburg Papers in Theoretical Linguistics* 76.

Anderson, A. H. (1995). 'Negotiating coherence in dialogue', in M. A. Gernsbacher and T. Givón (eds.), *Coherence in Spontaneous Text.* Amsterdam: John Benjamins.

——, Bader, M., Bard, E., Boyle, E., Doherty, G. M., and Garrod, S. (1991). 'The HCRC map task corpus', *Language and Speech* 34: 351–66.

André, E., Rist, T., and Müller, J. (1998). 'WebPersona: a life-like presentation agent for the world-wide web', *Knowledge-Based Systems* 11(1): 25–36.

Asher, N. and Lascarides, A. (1998). 'Questions in dialogue', *Linguistics and Philosophy* 21: 237–309.

Bangerter, A. (2004). 'Using pointing and describing to achieve joint focus of attention in dialogue', *Psychological Science* 15: 415–19.

—— and Clark, H. H. (2003). 'Navigating joint projects with dialogue', *Cognitive Science* 27: 195–225.

Bard, E. G. and Aylett, M. P. (2004). 'Referential form, word duration, and modelling the listener in spoken dialogue', in J. Trueswell and M. Tanenhaus (eds.), *Approaches to Studying World-situated Language Use*. Cambridge, MA: MIT Press.

Barr, D. J., and Keysar, B. (2002). Anchoring comprehension in linguistic precedents. *Journal of Memory and Language* 46: 391–418.

Bavelas, J. B. and Chovil, N. (2000). 'Visible acts of meaning. An integrated message model of language in face-to-face dialogue', *Journal of Language and Social Psychology* 19(2): 163–94.

——, Coates, L. and Johnson, T. (2000). 'Listeners as co-narrators', *Journal of Personality and Social Psychology* 79: 941–52.

Biederman, I. (1987). 'Recognition-by-components: a theory of human image understanding', *Psychological Review* 94(2): 115–47.

Bloom, P., Peterson, M. A., Nadel, L., and Garrett, M. (1996). *Language and Space.* Cambridge, MA: MIT Press.

Bock, K. (1989). 'Closed-class immanence in sentence production', *Cognition* 31: 163–86.

——, Dell, G. S., Chang, F., and Onishi, K. H. (2007). 'Persistent structural priming from language comprehension to language production', *Cognition* 104: 437–58.

Bolt, R. (1980). ' "Put-that-there": voice and gesture at the graphics interface', in *Proceedings of the Seventh Annual Conference on Computer Graphics and Interactive Techniques*: 262–70.

Branigan, H. P., Pickering, M. J., and Cleland, A. A. (2000). 'Syntactic co-ordination in dialogue', *Cognition* 75: B13–B25.

Brennan, S. E. and Clark, H. H. (1996). 'Conceptual pacts and lexical choice in conversation', *Journal of Experimental Psychology: Learning, Memory and Cognition* 22: 1482–93.

—— and Williams, M. (1995). 'The feeling of another's knowing: prosody and filled pauses as cues to listeners about the metacognitive states of speakers', *Journal of Memory and Language* 34: 383–98.

Bryant, D. J., Tversky, B., and Franklin, N. (1992). 'Internal and external spatial frameworks for representing described scenes', *Journal of Memory and Language* 31(1): 74–98.

—— and Wright, W. G. (1999). 'How bodily asymmetries determine accessibility in spatial frameworks', *Quarterly Journal of Experimental Psychology* 52A: 487–508.

Burke, C., Doran, C., Gertner, A., Gregorowicz, A., Harper, L., Korb, J., and Loehr, D. (2003). 'Dialogue complexity with portability? Research directions for the Information State approach', in *The Research Directions in Dialogue Processing Workshop at the 2003 HLT-NAACL/NSF Human Language Technology Conference.*

Bürkle, B., Nirmaier, H., and Herrmann, T. (1986). ' "Von dir aus ... ": zur hörerbezogenen lokalen Referenz. ["From your point of view ... ": on listener-oriented local reference]', Bericht Nr. 10, Arbeiten der Forschergruppe 'Sprechen und Sprachverstehen im sozialen Kontext', Heidelberg: Mannheim.

Cangelosi, A. and Parisi, D. (2002). 'Computer simulation: a new scientific approach to the study of language evolution', in A. Cangelosi and D. Parisi (eds.), *Simulating the Evolution of Language*. Berlin, Heidelberg, New York: Springer.

Carletta, J. C. (1996). 'Assessing agreement on classification tasks: the kappa statistic', *Computational Linguistics* 22(2): 249–54.

——, Isard, A., Isard, S., Kowtko, J. C., Doherty-Sneddon, G., and Anderson, A. H. (1997). 'The reliability of a dialogue structure coding schema', *Computational Linguistics*, 23(1), 13–31.

Carlson, L. A. (2000). 'Object use and object location: the effect of function on spatial relations', in E. van der Zee and U. Nikanne (eds.), *Cognitive Interfaces: Constraints on Linking Cognitive Information*. Oxford: Oxford University Press.

—— (2003). 'Using spatial language', *The Psychology of Learning and Motivation* 43: 127–61.

—— and Hill, P. L.. 'Formulating spatial descriptions across various dialogue contexts'.

—— —— (2008). 'Processing the presence, placement and properties of a distractor in spatial language tasks', *Memory and Cognition*, 36(2), 240–55.

—— and Logan, G. D. (2005). 'Attention and spatial language', in L. Itti, G. Rees, and J. Tsotsos (eds.). *Neurobiology of Attention*. Amsterdam: Elsevier/Academic Press.

—— and van der Zee, E. (eds.) (2005). *Functional Features in Language and Space: Insights from Perception, Categorization and Development*. Oxford: Oxford University Press.

Carlson-Radvansky, L. A. and Irwin, D. E. (1993). Frame of reference in vision and language: Where is above? *Cognition* 46(3): 223–244.

—— and Jiang, Y. (1998). 'Inhibition accompanies reference-frame selection', *Psychological Science* 9: 386–91.

—— and Logan, G. D. (1997). 'The influence of reference frame selection on spatial template construction', *Journal of Memory and Language* 37: 411–37.

—— and Radvansky, G. A. (1996). 'The influence of functional relations on spatial term selection', *Psychological Science* 7: 56–60.

—— and Tang, Z. (2000). 'Functional influences on orienting a reference frame', *Memory and Cognition* 28: 812–20.

Carroll, M. (1997). 'Changing place in English and German: language-specific preferences in the conceptualization of spatial relations', in J. Nuyts and E. Pederson (eds.), *Language and Conceptualization*. Cambridge: Cambridge University Press.

Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K., Tversky, D., Vaucelle, C., and Vilhjálmsson, H. (2002). 'MACK: Media lab Autonomous Conversational Kiosk', *Proceedings of Imagina '02*. 12–15 February, Monte Carlo.

Chan, T.-T. and Bergen, B. (2005). Writing Direction Influences Spatial Cognition. In B.G. Bara, L. Barsalou and M. Buchiarelli (eds.), *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Chartrand, T. L. and Bargh, J. A. (1999). 'The chameleon effect: the perception-behaviour link and social interaction', *Journal of Personality and Social Psychology* 76: 893–910.

Cheshire, J. (1996). 'That jacksprat: an interactional perspective on English that', *Journal of Pragmatics* 25: 369–93.

Clark, H. H. (1973). 'Space, time, semantics, and the child', in T. E. Moore (ed.), *Cognitive Development and the Acquisition of Language*. New York: Academic Press.

—— (1996). *Using Language.* Cambridge: Cambridge University Press.

—— and Wilkes-Gibbs, D. (1986). 'Referring as a collaborative process', *Cognition* 22: 1–39.

Cleland, A. A. and Pickering, M. J. (2003). 'The use of lexical and syntactic information in language production: evidence from the priming of noun phrase structure', *Journal of Memory and Language* 49: 214–30.

Clementini, E. and Di Felice, P. (1997). 'A global framework for qualitative shape description', *GeoInformatics* 1: 11–27.

Cohn, A. G., Bennett, B., Gooday, J., and Gotts, N. M. (1997). 'Qualitative spatial representation and reasoning with the region connection calculus', *Geoinformatics* 1: 1–44.

—— and Hazarika, S. (2001). 'Qualitative spatial representation and reasoning', *Fundamenta Informatica* 46: 1–29.

Conrad, F. G. and Schober, M. F. (2000). 'Clarifying question meaning in a household telephone survey', *Public Opinion Quarterly* 64: 1–28.

Core, M. and Allen, J. (1997). 'Coding dialogs with the DAMSL annotation scheme', in *Working Notes of the AAAI Fall Symposium on Communicative Actions in Humans and Machines*: 28–35, Cambridge, MA.

Cornish, F. (2001). 'Modal "that" as determiner and pronoun: the primacy of the cognitive-interactive dimension', *English Language and Linguistics* 5: 297–315.

Coventry, K. R. (1999). 'Function, geometry, and spatial prepositions: three experiments', *Spatial Cognition and Computation* 1: 145–54.

—— and Garrod, S. C. (2004). *Saying, Seeing and Acting. The Psychological Semantics of Spatial Prepositions*. Hove, UK: Psychology Press.

Dale, R., Geldof, S., and Prost, J.-P. (2005). 'Using natural language generation in automatic route description', *Journal of Research and Practice in Information Technology* 37(1): 89–105.

—— and Reiter, E. (1995). 'Computational interpretations of the Gricean maxims in the generation of referring expressions', *Cognitive Science* 18: 233–63.

Daniel, M. P., Tversky, B., and Heiser, J. (2006). 'How to put things together', manuscript.
de Vega, M., Rodrigo, M. J., Ato, M., Dehn, D. M., and Barquero, B. (2002). 'How nouns and prepositions fit together: an exploration of the semantics of locative sentences', *Discourse Processes* 34: 117–43.
Denis, M. (1997). 'The description of routes: a cognitive approach to the production of spatial discourse', *Cahiers de Psychologie Cognitive* 16: 409–58.
—— and Briault, X. (1997). 'Les aides verbales à la navigation', in M. Denis (ed.), *Langage et cognition spatiale, Sciences Cognitives*. Paris: Masson.
—— Pazzaglia, F., Cornoldi, C., and Bertolo, L. (1999). 'Spatial discourse and navigation: an analysis of route directions in the city of Venice', *Applied Cognitive Psychology* 13: 145–74.
Deutsch, W. and Pechmann, T. (1982). 'Social interactions and the development of definite descriptions', *Cognition* 11(2): 159–84.
Ehrich, V. (1985). 'Zur Linguistik und Psycholinguistik der sekundären Raumdeixis [Linguistics and psycholinguistics of secondary spatial deixis]', in H. Schweizer (ed.), *Sprache und Raum. Psychologische und linguistische Aspekte der Aneignung und Verarbeitung von Räumlichkeit. Ein Arbeitsbuch für das Lehren von Forschung*. Stuttgart: Metzler.
—— and Koster, C. (1983). 'Discourse organization and sentence form: the structure of room descriptions in Dutch', *Discourse Processes* 6: 169–95.
Eiser, J. R. (1971). 'Categorization, cognitive consistency and the concept of dimensional salience', *European Journal of Social Psychology* 1: 435–54.
Emmorey, K., Tversky, B., and Taylor, H. A. (2000). 'Using space to describe space: perspective in speech, sign, and gesture', *Spatial Cognition and Computation* 2: 157–80.
Enfield, N. J. (2004). 'On linear segmentation and combinatorics in co-speech gesture', *Semiotica*, 149-1/4, 57–123.
Engelkamp, J. (1998). *Memory for Actions*. Hove, UK: Psychology Press.
Engle, R. A. (1998). 'Not channels but composite signals: speech, gesture, diagrams, and object demonstrations are integrated in multimodal explanations', in M. A. Gernsbacher and S. J. Derry (eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.
Eschenbach, C. (2005). 'Contextual, functional, and geometric components in the semantics of projective terms', in L. Carlson and E. van der Zee (eds.), *Functional Features in Language and Space: Insights from Perception, Categorization, and Development*. Oxford: Oxford University Press.
Evers, M., Theune, M., and Karreman, J. (2007). 'Which way to turn? Guide orientation in virtual way finding', in *Proceedings of the ACL 2007 Workshop on Embodied Language Processing*: pp. 25–32.
Farrell, W. S. (1979). 'Coding left and right', *Journal of Experimental Psychology: Human Perception and Performance* 5: 42–51.
Fetzer, A. and Fischer, K. (eds). (2007). *Lexical Markers of Common Ground*. Elsevier.
Filipi, A. and Wales, R. (2003). 'Differential uses of okay, right and alright and their function in signaling perspective shift or maintenance in a map task', *Semiotica* 1: 429–55.
—— (2004). 'Perspective-taking and perspective-shifting as socially situated and collaborative actions', *Journal of Pragmatics* 36(10): 1851–84.
Fillmore, C. J. (1971). 'Toward a theory of deixis', *Working Papers in Linguistics* 3. Honolulu: University of Hawaii.

Fischer, K. (2003). 'Linguistic methods for investigating concepts in use', in T. Stolz and K. Kolbe (eds.), *Methodologie in der Linguistik*. Frankfurt a.M.: Lang.

—— and Moratz, R. (2001). 'From communicative strategies to cognitive modelling', in *Workshop Epigenetic Robotics*, Lund.

Formal Systems (2001). *Failures Divergence Refinement FDR2 Preliminary Manual*. Formal Systems (Europe) Ltd.

Foster, M. E. (2004). 'Corpus-based planning of deictic gestures in COMIC', in A. Belz, R. Evans, and P. Piwek (eds.), *Proceedings of the Third International Conference on Natural Language Generation*: 198–204, Springer, Lecture Notes in Computer Science, volume 3123.

Franklin, N., Henkel, L. A., and Zangas, T. (1995). 'Parsing surrounding space into regions', *Memory and Cognition* 23: 397–407.

—— and Tversky, B. (1990). 'Searching imagined environments', *Journal of Experimental Psychology* 119: 63–76.

Freksa, C. (1992). 'Using orientation information for qualitative spatial reasoning', in A. U. Frank, I. Campari, and U. Formentini (eds.), *Theories and Methods of Spatio-temporal Reasoning in Geographic Space*. Berlin, Heidelberg: Springer.

Gardner, R. (2001). *When Listeners Talk. Pragmatics and Beyond*. John Benjamins.

Garrod, S. C. and Anderson, A. (1987). 'Saying what you mean in dialogue: a study in conceptual and semantic co-ordination', *Cognition* 27: 181–218.

—— and Doherty, G. (1994). 'Conversation, co-ordination and convention: an empirical investigation of how groups establish linguistic conventions', *Cognition* 53: 181–215.

—— and Pickering, M. J. (2004). 'Why is conversation so easy?', *Trends in Cognitive Science* 8: 8–11.

—— —— (2007). 'Alignment in dialogue', in G. Gaskell (ed.), *Oxford Handbook of Psycholinguistics*. Oxford: Oxford University Press.

Giles, H. and Powesland, P. F. (1975). *Speech Styles and Social Evaluation*. New York: Academic Press.

Ginzburg, J. (1996). 'Interrogatives: questions, facts and dialogue', in S. Lappin (ed.), *Handbook of Contemporary Semantic Theory*. Oxford: Blackwell.

Glover, K. (2000). 'Proximal and distal deixis in negotiation talk', *Journal of Pragmatics* 32: 915–26.

—— and Grundy P. (1996). 'Why do we have these: when reconstructing the indexical ground is disfavoured'. *Time, Space and Identity. Second International Colloquium on Deixis*. Nancy: Centre de Recherche en Informatique de Nancy, 117–33.

Goel, V. (1995). *Sketches of Thought*. Cambridge: MIT Press.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Goldin-Meadow, S. (2003). *Hearing Gesture: How our Hands Help us Think*. Cambridge: Belknap Press.

Golding, J., Graesser, A., and Hauselt, J. (1996). 'The process of answering direction-giving questions when someone is lost on a university campus: the role of pragmatics', *Applied Cognitive Psychology* 10: 23–39.

Goldschmidt, G. (1991). 'The dialectics of sketching', *Creativity Research Journal* 4, 123–43.

Gorniak, P. and Roy, D. (2004). 'Grounded semantic composition for visual scenes', *Journal of Artificial Intelligence Research* 21: 429–70.

Grice, H. P. (1975). 'Logic and conversation (from the William James lectures, Harvard University, 1967)', in P. Cole and J. Morgan (eds.), *Syntax and Semantics 3: Speech Acts*. New York: Academic Press.

Gries, S. T. (2005). 'Syntactic priming: a corpus-based approach', *Journal of Psycholinguistic Research* 34: 365–99.

Grosz, B. J. and Sidner, C. L. (1986). 'Attention, intentions, and the structure of discourse', *Computational Linguistics* 12(3): 175–204.

Habel, C. (2003). 'Incremental generation of multimodal route instructions', in *Proceedings of the AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue*.

Hanks, W. F. (1990). *Referential Practice: Language and Lived Space among the Mayas*. Chicago: University of Chicago Press.

—— (1992). 'The indexical ground of deictic reference', in A. Duranti and C. Goodwin (eds.), *Rethinking Context: Language as an Interactional Phenomenon*. New York: Cambridge University Press.

Harnad, S. (1987). *Categorical Perception: The Groundwork of Cognition*. Cambridge: Cambridge University Press.

Hartsuiker, R. J., Pickering, M. J., and Veltkamp, E. (2004). 'Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals', *Psychological Science* 15: 409–14.

Haubensak, G. (1992). 'The consistency model: a process model for absolute judgments', *Journal of Experimental Psychology: Human Perception and Performance* 18: 303–9.

Hayward, W. G. and Tarr, M. J. (1995). 'Spatial language and spatial representation', *Cognition* 55: 39–84.

Haywood, S. L., Pickering, M. J., and Branigan, H. P. (2005). 'Do speakers avoid ambiguities during dialogue?', *Psychological Science* 16: 362–6.

Hegarty, M. and Waller, D. (2004). 'A dissociation between mental rotation and perspective-taking abilities', *Intelligence* 32: 175–91.

Heiser, J., Phan, D., Agrawala, M., Tversky, B., and Hanrahan, P. (2004). 'Identification and validation of cognitive design principles for automated generation of assembly instructions', *Proceedings of Advanced Visual Interfaces '04* ACM: 311–19.

Herrmann, T. (1989). 'Partnerbezogene Objektlokalisation—ein neues sprachpsychologisches Forschungsthema. [Partner-oriented localization of objects—a new psycholinguistic research topic]', Bericht Nr. 25, Arbeiten der Forschergruppe 'Sprechen und Sprachverstehen im sozialen Kontext', Heidelberg: Mannheim.

—— (1990). 'Vor, hinter, rechts und links: das 6H-Modell [In front, behind, right and left: the 6H model]', *Zeitschrift für Literaturwissenschaft und Linguistik* 78: 117–40.

—— Bürkle, B., and Nirmaier, H. (1987). 'Zur hörerbezogenen Raumreferenz: Hörerposition und Lokalisationsaufwand [On listener-oriented spatial reference: Listener position and localization effort]', *Sprache & Kognition* 3: 126–37.

—— and Deutsch, W. (1976). *Psychologie der Objektbenennung*. Bern: Huber.

—— and Grabowski, J. (1994). *Sprechen. Psychologie der Sprachproduktion*. Heidelberg: Spektrum.

—— and Schweizer, K. (1998). *Sprechen über Raum. Sprachliches Lokalisieren und seine kognitiven Grundlagen* [Speaking about space. Verbal localization and its cognitive bases]. Bern: Huber.

Herskovits, A. (1986). *Language and Spatial Cognition: an Interdisciplinary Study of the Prepositions of English*. Cambridge: Cambridge University Press.

Hindmarsh, J. and Heath, C. (2000). 'Embodied reference: a study of deixis in workplace interaction', *Journal of Pragmatics* 32: 1855–78.

Hoare, C. A. R. (1985). *Communicating Sequential Processes*. Prentice-Hall.

Holzapfel, H. and Gieselmann, P. (2004). 'A way out of dead end situations in dialogue systems for human-robot interaction', in *Humanoids 2004*. Los Angeles.

Horton, W. S. and Gerrig, R. J. (2005). 'The impact of memory demands on audience design during language production', *Cognition* 96: 127–42.

—— and Keysar, B. (1996). 'When do speakers take into account common ground?', *Cognition* 59: 91–117.

Hummels, C. (2000). *Gestural design tools: prototypes, experiments and scenarios.* Doctoral dissertation, Technische Universiteit Delft.

Iachini, T. and Logie, R. H. (2003). 'The role of perspective in locating position in a real-world, unfamiliar environment', *Applied Cognitive Psychology* 17(6): 715–32.

Isaacs, E. A. and Clark, H. H. (1987). 'References in conversation between experts and novices', *Journal of Experimental Psychology: General* 116(1): 26–37.

Jackendoff, R. (1996). 'The architecture of the linguistic-spatial interface', in P. Bloom, M. A. Peterson, L. Nadel, and M. F. Garrett (eds.), *Language and Space: Language, Speech and Communication*. Cambridge, MA: MIT Press.

Johnston, M. and Bangalore, S. (2000). 'Finite state multimodal parsing and understanding', in *Proceedings of the International Conference on Computational Linguistics (COLING, 2000)*: 369–75.

Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., and Ess-Dykema, C. V. (1997). 'Switchboard discourse language modeling project final report', Summer Research Workshop Technical Reports, Research Note 30, Johns Hopkins University, Baltimore, MD.

Just, M. A. and Carpenter, P. A. (1985). 'Cognitive coordinate systems: accounts of mental rotation and individual differences in spatial ability', *Psychological Review* 92: 137–72.

Kataoka, K. (2004). 'Co-construction of a mental map in spatial discourse: a case study of Japanese rock climbers' use of deictic verbs of motion', *Pragmatics* 14: 409–38.

Kelly, S. D., Kravitz, C., and Hopkins, M. (2004). 'Neural correlates of bimodal speech and gesture comprehension', *Brain and Language* 89(1): 253–60.

Kendon, A. (1988). *Sign Languages of Aboriginal Australia: Cultural, Semiotic and Communicative Perspectives*. Cambridge: Cambridge University Press.

—— (2004). *Gesture. Visible action as utterance.* Cambridge, UK: Cambridge University Press.

Kessell, A. and Tversky, B. (2005). 'Gestures for thinking and explaining', in *Proceedings of the 27*th *Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

Kimbara, I. (2006). 'On gestural mimicry', *Gesture* 6(1): 39–61.

Klein, W. (1982). 'Local deixis in route directions', in R. Jarvella and W. Klein (eds.), *Speech, Place and Action*. Chichester, UK: Wiley.

Klippel, A., Tenbrink, T., and Montello, D. (in press). 'The role of structure and function in the conceptualization of directions', in M. Dimitrova-Vulchanova and E. van der Zee (eds.), *Motion Encoding in Spatial Language*. Oxford: Oxford University Press.

Koons, D. B., Sparrell, C. J., and Thorisson, K. R. (1993). 'Integrating simultaneous input from speech, gaze and hand gestures', in M. T. Maybury (ed.), *Intelligent Multimedia Interfaces*. Cambridge, MA: MIT Press.

Kopp, S., Tepper, P., and Cassell, J. (2004). 'Towards integrated microplanning of language and iconic gesture for multimodal output', in *Proceedings of the Sixth International Conference on Multimodal Interfaces*. New York: ACM Press.

—— ——, Ferriman, K., Striegnitz, K., and Cassell, J. (2007). 'Trading spaces: how humans and humanoids use speech and gesture to give directions', in T. Nishida (ed.) *Conversational Informatics: An Engineering Approach*. John Wiley and Sons.

—— and Wachsmuth, I. (2004). 'Synthesizing multimodal utterances for conversational agents', *The Journal of Computer Animation and Virtual Worlds* 15(1): 39–52.

Kosko, B. (1988). 'Bidirectional associative memories' *IEEE Transactions on Systems, Man and Cybernetics* 18(1): 49–60.

Kowtko, J. (1996). 'The function of intonation in task-oriented dialogues', PhD thesis, University of Edinburgh.

Krauss, R. M. (1998). 'Why do we gesture when we speak?', *Current Directions in Psychological Science* 7: 54–9.

—— Dushay, R. A., Chen, Y., and Rauscher, F. (1995). 'The communicative value of conversational hand gestures', *Journal of Experimental Social Psychology* 31: 533–52.

Krieg-Brückner, B., Frese, U., Lüttich, K., Mandel, C., Mossakowski, T., and Ross, R. (2005). 'Specification of an ontology for route graphs', in C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel, and T. Barkowsky (eds.), *Spatial Cognition IV: Reasoning, Action, Interaction*. Berlin, Heidelberg: Springer.

—— and Shi, H. (2006). 'Orientation calculi and route graphs: towards semantic representations for route descriptions', in M. Raubal, H. J. Miller, A. U. Frank, and M. F. Goodchild (eds.) *Proceedings International Conference GIScience 2006, Münster, Germany*. Berlin, Heidelberg: Springer.

Kronmüller, E. and Barr, D. J. (2007). 'Perspective-free pragmatics: broken precedents and the recovery-from-preemption hypothesis', *Journal of Memory and Language* 56: 436–55.

Kruijff, G.-J. M., Zender, H., Jensfelt, P., and Christensen, H. I. (2007). 'Situated dialogue and spatial organization: what, where... and why?', *International Journal of Advanced Robotic Systems, Special Issue on Human and Robot Interactive Communication* 4(2).

Kurtz, V. and Schober, M. F. (2001). 'Readers' varying interpretations of theme in short fiction', *Poetics* 29: 139–66.

Kyriacou T., Bugmann G., and Lauria S. (2005). 'Vision-based urban navigation procedures for verbally instructed robots', *Robotics and Autonomous Systems* 51: 69–80.

Lakoff, G. (1972). 'Hedges: a study in meaning criteria and the logic of fuzzy concepts', *Journal of Philosophical Logic* 2: 458–508.

Landau, B. (1996). 'Multiple geometric representations of objects in languages and language learners', in P. Bloom, M. A. Peterson, L. Nadel, and M. Garrett (eds.), *Language and Space: Language, Speech and Communication*. Cambridge, MA: MIT Press.

—— and Jackendoff, R. (1993). ' "What" and "where" in spatial language and spatial cognition.' *Behavioural and Brain Sciences* 16: 217–65.

Lang, E. (1989). 'The semantics of dimensional designation of spatial objects', in M. Bierwisch and E. Lang (eds.), *Dimensional Adjectives: Grammatical Structure and Conceptual Interpretation*. Berlin, Heidelberg, New York: Springer.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar, Volume 1*. Stanford, CA: Stanford Press.

Lankenau, A. and Röfer, T. (2000). 'The role of shared control in service robots—the Bremen autonomous wheelchair as an example', in T. Röfer, A. Lankenau, and R. Moratz (eds.), *Service Robotics—Applications and Safety Issues in an Emerging Market. Workshop Notes. European Conference on Artificial Intelligence 2000* (ECAI 2000).

Larsson, S. (2002). 'Issue-based dialogue management', PhD thesis, Göteborg University.

—— and Traum, D. (2000). 'Information state and dialogue management in the TRINDI dialogue move engine toolkit', in *Natural Language Engineering, Special Issue on Best Practice in Spoken Language Dialogue Systems Engineering*.

Lemon, O., Bracy, A., Gruenstein, A., and Peters, S. (2003). An Information State Approach in a Multi-modal Dialogue System for Human-Robot Conversation, in P. Kuhnlein, H. Reiser, and H. Zeevat (eds.), *Perspectives on Dialogue in the New Millennium*. John Benjamins Publishing Company.

Levelt, W. J. M. (1982). 'Cognitive styles in the use of spatial direction terms', in R. Jarvella and W. Klein (eds.), *Speech, Place, and Action*. Chichester, UK: Wiley.

—— (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Levinson, S. C. (1996). Frames of reference and Molyneux's question: cross-linguistic evidence', in P. Bloom, M. A. Peterson, L. Nadel, and M. Garrett (eds.), *Language and Space: Language, Speech and Communication*. Cambridge, MA: MIT Press.

—— (2003). *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge: Cambridge University Press.

Ligozat, G. (2000). 'From language to motion, and back: generating and using route descriptions', in D. N. Christodoulakis (ed.), *NLP 2000, LNCS 1835* (pp. 328–45). Berlin, Heidelberg: Springer.

Loetzsch, M., Risler, M., and Jüngel, M. (2006). 'XABSL—A pragmatic approach to behavior engineering', in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2006)*. Beijing.

Logan, G. D. (1994). 'Spatial attention and the apprehension of spatial relations', *Journal of Experimental Psychology: Human Perception and Performance* 20: 1015–36.

—— (1995). 'Linguistic and conceptual control of visual spatial attention', *Cognitive Psychology* 28: 103–74.

—— (1996). 'The CODE theory of visual attention: an integration of space-based and object-based attention', *Psychological Review* 103: 603–49.

—— and Sadler (1996). 'A computational analysis of the apprehension of spatial relations', in P. Bloom, M. A. Peterson, L. Nadel, and M. Garrett (eds.), *Language and Space: Language, Speech and Communication*. Cambridge, MA: MIT Press.

Look, G., Kottahachchi, B., Laddaga, R., and Shrobe, H. (2005). 'A location representation for generating descriptive walking directions', in *IUI '05: Proceedings of the Tenth International Conference on Intelligent User Interfaces*.

Mainwaring, S. D., Tversky, B., Ohgishi, M., and Schiano, D. J. (2003). 'Descriptions of simple spatial scenes in English and Japanese', *Spatial Cognition and Computation* 3(1): 3–42.

Maki, R. H., Grandy, C. A., and Hauge, G. (1979). 'Why is telling right from left more difficult than telling above from below?', *Journal of Experimental Psychology: Human Perception and Performance* 5: 52–67.

Mangold, R. and Pobel, R. (1988). 'Informativeness and instrumentality in referential communication', *Journal of Language and Social Psychology* 7(3–4): 181–91.

Marr, D. and Nishihara, H. (1978). 'Representation and recognition of the spatial organization of three-dimensional shapes', *Proceedings of the Royal Society, Series B* 200: 269–94.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought.* Chicago: University of Chicago Press.

—— (2005). *Gesture & Thought*. Chicago: University of Chicago Press.

—— and Levy, E. (1982). 'Conceptual representations in language activity and gesture', in R. Jarvella and W. Klein (eds.), *Speech, Place, and Action* (pp. 271–95). Chichester, UK: Wiley.

Metzing, C. and Brennan, S. E. (2003). 'When conceptual pacts are broken: partner-specific effects on the comprehension of referring expressions', *Journal of Memory and Language* 49: 201–13.

Michon, P. E. and Denis, M. (2001). 'When and why are visual landmarks used in giving directions?', in D. R. Montello (ed.), *Spatial Information Theory: Foundations of Geographic Information Science*. COSIT 2001 Morro Bay, CA, USA, 19–23 September, 2001. *Proceedings*. Berlin, Heidelberg: Springer.

Millar, B., Vonwiller, J., Harrington, J., and Dermody, P. (1994). 'The Australian national database of spoken language', *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 94/1: 97–100. Adelaide.

Miller, G., and Johnson, P. (1976). *Language and Perception.* Cambridge MA: Harvard University Press.

Miyake, A. (2001). 'Individual differences in WM: introduction to the special section', *Journal of Experimental Psychology: General* 130: 163–8.

Molder, H. T. and Potter, J. (eds.) (2005). *Conversation and Cognition*. Cambridge: Cambridge University Press.

Moratz, R. and Tenbrink, T. (2006). 'Spatial reference in linguistic human-robot interaction: iterative, empirically supported development of a model of projective relations', *Spatial Cognition and Computation* 6(1): 63–106.

—— ——, Bateman, J., and Fischer, K. (2003). 'Spatial knowledge representation for human-robot interaction', in C. Freksa, W. Brauer, C. Habel, and K. F. Wender (eds). *Spatial Cognition III*. Berlin: Springer.

Mortenson, M. E. (1997). *Geometric Modeling*. New York: Wiley.

Muller, P. and Prévot, L.. 'Grounding information in route explanation dialogues'.

Novick, L. R. and Morse, D. L. (2000). 'Folding a fish, making a mushroom: the role of diagrams in executing assembly procedures', *Memory and Cognition* 28: 1242–56.

Ochs, E., Schegloff, E. A., and Thompson, S. A. (eds.) (1996). *Interaction and Grammar*. Cambridge: Cambridge University Press.

Olson, D. R. (1970). 'Language and thought: aspects of a cognitive theory of semantics', *Psychological Review 77*: 257–73.

Östman, J. O. (1995). 'Recasting the deictic foundation, using physics and Finnish', in M. Shibatani and S. Thompson (eds.), *Essays in Semantics and Pragmatics in Honor of Charles J. Fillmore*. Amsterdam and Philadelphia: John Benjamins.

Paivio, A. (1986). *Mental Representations*. New York: Oxford University Press.

Pickering, M. J. and Garrod, S. C. (2004). 'Towards a mechanistic psychology of dialogue', *Behavioural and Brain Sciences* 27(2): 169–90.

—— —— (2006). 'Alignment as the basis for successful communication'. *Research on Language and Computation* 4: 203–28.

Prévot, L. (2004). 'Structures sémantiques et pragmatiques pour la modélisation de la cohérence dans des dialogues nalises', PhD thesis, Université Paul Sabatier.

Psathas, G. (1986). 'Some sequential structures in direction-giving', *Human Studies* 9: 231–46.

—— (1991). 'The structure of direction-giving in interaction', in D. Boden and D. H. Zimmerman (eds.), *Talk and Social Structure: Studies in Ethnomethodology and Conversation Analysis*. Cambridge: Polity Press.

Pylyshyn, Z. (2002). 'Mental imagery: in search of a theory', *Behavioral and Brain Sciences* 25(2): 157–237.

Regier, T. and Carlson, L. A. (2001). 'Grounding spatial language in perception. An empirical and computational investigation', *Journal of Experimental Psychology: General* 130: 273–98.

Reips, U.-D. (2002). 'Theory and techniques of Web experimenting', in B. Batinic, U.-D. Reips, and M. Bosnjak (eds.), *Online Social Sciences*. Seattle: Hogrefe and Huber.

Riesbeck, C. K. (1980). 'You can't miss it! Judging clarity of directions', *Cognitive Science* 4: 285–303.

Rimé, B. and Schiaratura, L. (1991). 'Gesture and speech', in R. S. Feldman and B. Rimé (eds.), *Fundamentals of Nonverbal Behavior*. Cambridge: Cambridge University Press.

Röfer, T., Brunn, R., Dahm, I., Hebbel, M., Hoffmann, J., Jungel, M., et al. (2004). 'GermanTeam 2004', in *Robocup 2004: Robot Soccer World Cup VIII Preproceedings. Lisbon, Portugal: RoboCup Federation*. (Extended version (299 pages) at http://www.germanteam.org/GT2004.pdf).

Roscoe, A. W. (1998). *The Theory and Practice of Concurrency*. Prentice-Hall.

Ross, R., Bateman, J., and Shi, H. (2005). 'Using generalized dialogue models to constrain information state based dialogue systems', in *Proceedings of the Symposium on Dialogue Modelling and Generation.* http://lubitsch.lili.uni-bielefeld.de/ DMG/Proceedings/proc.html.

—— Shi, H., Vierhuff, T., Krieg-Brückner, B., and Bateman, J. (2005). 'Towards dialogue-based shared control of navigating robots', in C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel, and T. Barkowsky (eds.), *Spatial Cognition IV: Reasoning, Action, Interaction*. Berlin, Heidelberg: Springer.

Rossari, C., Beaulieu-Masson, A., Cojocariu, C., and Razgouliaeva, A. (2004). *Autour des connecteurs. Réflexions sur l'énonciation et la portée*, volume 75 of Sciences pour la communication. Bern: Lang.

Roulet, E., Auchlin, A., Moeschler, J., Schelling, M., and Rubattel, C. (1985). *L'articulation du discours en français contemporain*. (Collection Sciences pour la communication). Bern: Lang.

Russell, A. W., and Schober, M. F. (1999). 'How beliefs about a partner's goals affect referring in goal-discrepant conversations', *Discourse Processes* 27(1): 1–33.

Safarova, M., Muller, P., and Prévot, L. (2005). 'The discourse function of final rises in French dialogues', in C. Gardent and B. Gaie (eds.), *Ninth Workshop On The Semantics And Pragmatics Of Dialogue (SemDial)*.

Schegloff, E. A. (1972). 'Notes on a conversational practice: formulating place', in D. Sudnow (ed.), *Studies in Social Interaction*. New York: Free Press.

—— (1991). 'Conversation analysis and socially shared cognition', in L. Resnick, J. Levine, and S. Teasley (eds.), *Perspectives on Socially Shared Cognition*. Washington: APA.

Schegloff, E. A. (1992). 'Repair after next turn: the last structurally provided place for the defence of intersubjectivity in conversation', *American Journal of Sociology* 95: 1295–345.

Schiffrin, D. (1987). *Discourse Markers*. Cambridge: Cambridge University Press.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition* 47(1): 1–24.

—— (1995). 'Speakers, addressees, and frames of reference: whose effort is minimized in conversations about location?', *Discourse Processes* 20(2): 219–47.

—— (1998a). 'Different kinds of conversational perspective-taking', in S. R. Fussell and R. J. Kreuz (eds.), *Social and Cognitive Psychological Approaches to Interpersonal Communication*. Mahwah, NJ: Lawrence Erlbaum.

—— (1998b). 'How addressees affect spatial perspective choice in dialogue', in P. L. Olivier and K.-P. Gapp (eds.), *Representation and Processing of Spatial Expressions*. Mahwah, NJ: Lawrence Erlbaum.

—— (1998c). 'How partners with high and low spatial ability choose perspectives in conversation', *Abstracts of the 39th Annual Meeting of the Psychonomic Society*, Dallas, TX.

—— (2005). 'Conceptual alignment in conversation', in B. F. Malle and S. D. Hodges (eds.), *Other Minds: How Humans Bridge the Divide between Self and Others*. New York: Guilford Press.

—— (2006). 'Dialogue and interaction', in K. Brown (ed.), *The Encyclopedia of Language and Linguistics, 2nd Edition.* Oxford: Elsevier.

——. 'Spatial dialogue between partners with mismatched abilities'.

—— and Bloom, J. E. (2004). 'Discourse cues that respondents have misunderstood survey questions', *Discourse Processes* 38: 287–308.

—— and Brennan, S. E. (2003). 'Processes of interactive spoken discourse: the role of the partner', in A. C. Graesser, M. A. Gernsbacher, and S. R. Goldman (eds.), *Handbook of Discourse Processes*. Mahwah, NJ: Lawrence Erlbaum.

—— and Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, 21: 211–232.

—— Conrad, F. G., and Fricker, S. S. (2004). 'Misunderstanding standardized language in research interviews'. *Applied Cognitive Psychology* 18: 169–88.

Schon, D. A. (1983) *The Reflective Practitioner.* New York: Basic Books.

Schoonbaert, S., Hartsuiker, R. J., and Pickering, M. J. (2007). 'The representation of lexical and syntactic information in bilinguals: evidence from syntactic priming', *Journal of Memory and Language* 56: 153–71.

Shatz, M. and Gelman, R. (1973). 'The development of communication skills: modifications in the speech of young children as a function of listener', *Monographs of the Society for Research in Child Development* 38(5): 1–38.

Shi, H., Mandel, C., and Ross, R. J. (2007). 'Interpreting route instructions as qualitative spatial actions', in T. Barkowsky, M. Knauff, G. Ligozat, and D. Montello (eds.), *Spatial Cognition V: Reasoning, Action, Interaction*. Berlin: Springer.

—— Ross, R., and Bateman, J. (2005). 'Formalising control in robust spoken dialogue systems', in B. K. Aichernig and B. Beckert (eds.), *Proceedings of 3rd IEEE International Conference on Software Engineering and Formal Methods*.

Shockley, K., Santana, M. V., and Fowler, C. A. (2003). 'Mutual interpersonal postural constraints are involved in cooperative conversation', *Journal of Experimental Psychology: Human Perception and Performance* 29: 326–32.

Sitter, S. and Stein, A. (1992). 'Modelling the illocutionary aspects of information-seeking dialogues', *Information Processing and Management* 28: 124–35.

Skubic, M., Perzanowski, D., Blisard, S., Schultz, A., Adams, W., Bugajska, M., and Brock, D. (2004). 'Spatial language for human-robot dialogs', *IEEE Transactions on Systems, Man and Cybernetics, Part C* 34(2): 154–67.

Slobin, D. I. (1996). 'From "thought and language" to "thinking for speaking" ', in S. C. Levinson and J. J. Gumperz. (eds.), *Studies in the Social and Cultural Foundations of Language, No. 17*. New York, NY: Cambridge University Press.

Soto, D. and Blanco, M. J. (2004). 'Spatial attention and object-based attention: a comparison within a single task', *Vision Research* 44(1): 69–81.

Sowa, T. (2006a). *Understanding Coverbal Iconic Gestures in Shape Descriptions.* Berlin: Akademische Verlagsgesellschaft Aka.

—— (2006b). 'Towards the integration of shape-related information in 3-D gestures and speech', in *Proceedings of the Eighth International Conference on Multimodal Interfaces*. New York: ACM Press.

—— and Wachsmuth, I. (2002). 'Interpretation of shape-related iconic gestures in virtual environments', in I. Wachsmuth and T. Sowa (eds.), *Gesture and Sign Language in Human-Computer Interaction*. Berlin: Springer.

—— and Wachsmuth, I. (2003). 'Coverbal iconic gestures for object descriptions in virtual environments: an empirical study', in M. Rector, I. Poggi, and N. Trigo (eds.), *Gestures: Meaning and Use*. Porto, Portugal: Ediçôes Universidade Fernando Pessoa.

Sperber, D. and Wilson, D. (1995). *Relevance: Communication and Cognition*. Oxford: Blackwell.

Spexard, T., Li, S., Wrede, B., Fritsch, J., Sagerer, G., Booij, O., Zivkovic, Z., Terwijn, B., and Kröse, B. (2006). 'BIRON, where are you? Enabling a robot to learn new places in a real home environment by integrating spoken dialog and visual localization', in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Steels, L. (1996). 'Perceptually grounded meaning creation', in M. Tokoro (ed.), *Proceedings of the Second International Conference on Multi-Agent Systems ICMAS '96*. Menlo Park, CA: AAAI Press.

—— (2001). 'Language games for autonomous robots', *IEEE Intelligent Systems* 16(5): 16–22.

—— (2003). 'Evolving grounded communication for robots', *Trends in Cognitive Sciences* 7(7): 308–12.

—— and Belpaeme, T. (2005). 'Coordinating perceptually grounded categories through language: a case study for colour', *Behavioral and Brain Sciences* 28(4): 469–89.

—— and Brooks, R. (1994). *The Artificial Life Route to Artificial Intelligence. Building Situated Embodied Agents.* New Haven: Lawrence Erlbaum.

—— and De Beule, J. (2006). 'Unify and merge in fluid construction grammar', in P. Vogt, Y. Sugita, E. Tuci, and C. Nehaniv (eds.), *Symbol Grounding and Beyond: Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication, EELC 2006, Vol. 4211*. Berlin, Heidelberg, New York: Springer.

—— and M. Loetzsch. 'Perspective alignment in spatial language'.

Suwa, M., Tversky, B., Gero, J., and Purcell, T. (2001). 'Seeing into sketches: regrouping parts encourages new interpretations', in J. S. Gero, B. Tversky, and T. Purcell (eds.), *Visual and Spatial Reasoning in Design*. Sydney, Australia: Key Centre of Design Computing and Cognition.

Talmy, L. (1983). 'How language structures space', in H. L. Pick and L. P. Acredolo (eds.), *Spatial Orientation: Theory, Research, and Application*. New York: Plenum Press.

—— (1996). 'Fictive motion in language and "ception"', in P. Bloom, M. A. Peterson, L. Nadel, and M. Garrett (eds.), *Language and Space: Language, Speech and Communication*. Cambridge, MA: MIT Press.

—— (2000). *Towards a Cognitive Semantics*. Cambridge, MA: MIT Press.

Taylor, H. A. and Tversky, B. (1992a). 'Spatial mental models derived from survey and route descriptions', *Journal of Memory and Language* 31: 261–92.

—— —— (1992b). 'Descriptions and depictions of environments', *Memory and Cognition* 20: 483–96.

—— (1996). Perspective in spatial descriptions. *Journal of Memory and Language* 35: 371–91.

Tenbrink, T. (2005). 'Identifying objects on the basis of spatial contrast: an empirical study', in C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel, and T. Barkowsky (eds.), *Spatial Cognition IV: Reasoning, Action, Interaction*. Berlin, Heidelberg: Springer.

—— (2006). 'Teaching an autonomous wheelchair where things are', in K. Fischer (ed.), *Proceedings of the Workshop on 'How People Talk to Computers, Robots, and Other Artificial Communication Partners', Hansewissenschaftskolleg, Delmenhorst, 21–23 April 2006*. SFB/TR8 Report 010-09_2006.

—— (2007). *Space, Time, and the Use of Language: an Investigation of Relationships*. Berlin: Mouton de Gruyter.

—— and Shi, H. 2007. 'Negotiating spatial goals with a wheelchair', in S. Keizer, H. Bunt, and T. Paek (eds.), *Proceedings of the Eighth SIGdial Workshop on Discourse and Dialogue*. Antwerp, Belgium, 1–2 September 2007.

Testa, L. (2005). 'Referential communication with adults with mental retardation: staff accommodation', Unpublished doctoral dissertation, New School for Social Research.

Theune, M., Hofs, D., and van Kessel, M. (2007). 'The virtual guide: a direction-giving embodied conversational agent', in *Proceedings of Interspeech 2007*: 2197–200.

Thibault, P. J. (2006). *Brain, Mind and the Signifying Body: an Ecosocial Semiotic Theory*. London: Continuum.

Tomasello, M. (1995). 'Joint attention as social cognition', in C. Moore and P. J. Dunham (eds.), *Joint Attention: its Origins and Role in Development*. Hillsdale, NJ: Lawrence Erlbaum.

Tomko, M. and Winter, S. (2006). 'Recursive construction of granular route directions', *Journal of Spatial Science* 51(1): 101–15.

Traum, D. (1994). 'A computational theory of grounding in natural language conversation', PhD thesis, University of Rochester.

—— and Larsson S. (2003). 'The information state approach to dialogue management', in Smith and Kuppevelt (eds.), *Current and New Directions in Discourse and Dialogue*. Kluwer.

Tsal, Y. and Lavie, N. (1988). 'Attending to colour and shape: the special role of location in selective visual processing', *Perception and Psychophysics* 44: 15–21.

—— —— (1993). 'Location dominance in attending to colour and shape', *Journal of Experimental Psychology: Human Perception and Performance* 19: 131–9.

Tschander, L., Schmidtke, H., Habel, C., Eschenbach, C., and Kulik, L. (2003). 'A geometric agent following route instructions', in C. Freksa, W. Brauer, C. Habel, and K. F. Wender (eds.), *Spatial Cognition III*: *Routes and Navigation, Human Memory and Learning, Spatial Representation and Spatial Learning*. Springer.

Tulving E. (1972). 'Episodic and semantic memory', in E. Tulving and W. Donaldson (eds.), *Organization of Memory*. New York: Academic Press.

Tversky, B. (1996). 'Spatial perspective in descriptions', in P. Bloom, M. A. Peterson, L. Nadel, and M. F. Garrett (eds.), *Language and Space. Language, Speech and Communication*. Cambridge, MA: MIT Press.

—— (2001). 'Spatial schemas in depictions', in M. Gattis (ed.), *Spatial Schemas and Abstract Thought*. Cambridge, MA: MIT Press.

—— (2003). 'Structures of mental spaces—how people think about space', *Environment and Behavior* 35(1): 66–80.

—— Agrawala, M., Heiser, J., Lee, P. U., Hanrahan, P., Phan, D., Stolte, C., and Daniel, M.-P. (2007). 'Cognitive design principles for generating visualizations', in G. Allen (ed.), *Applied Spatial Cognition: from Research to Cognitive Technology*. Hillsdale, NJ: Lawrence Erlbaum.

—— Heiser, J., Lozano, S., MacKenzie, R., and Morrison, J. (2007). 'Enriching animations', in R. Lowe and W. Schnotz (eds.), *Learning with Animation.* Cambridge: Cambridge University Press.

—— and Lee, P. U. (1998). 'How space structures language', in C. Freksa, C. Habel, and K. F. Wender (eds.), *Spatial Cognition: an Interdisciplinary Approach to Representation and Processing of Spatial Knowledge*. Berlin: Springer.

—— —— (1999). 'Pictorial and verbal tools for conveying routes', in C. Freksa and D. M. Mark (eds.), *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science*. Berlin: Springer.

—— Zacks, J. M., and Hard, B. M. (2008). 'The structure of experience', in T. Shipley and J. M. Zacks (eds.), *Understanding Events*. Oxford: Oxford University Press.

—— —— Lee, P. U., and Heiser, J. (2000). 'Lines, blobs, crosses, and arrows: diagrammatic communication with schematic figures', in M. Anderson, P. Cheng, and V. Haarslev (eds.), *Theory and Application of Diagrams*. Berlin: Springer.

—— Zacks, J. M., and Martin, B. (2008). The structure of experience. In T. F. Shipley and J. M. Zacks (eds.), *Understanding events: From perception to action*. (pp. 436–464).

Ullmer-Ehrich, V. (1982). 'The structure of living space descriptions', in R. J. Jarvella and W. Klein (eds.), *Speech, Place, and Action*. Chichester: Wiley.

van Berkum, J. J. A., van den Brink, D., Tesink, C., Kos, M., and Hagoort, P. (2008). 'The neural integration of speaker and message', *Journal of Cognitive Neuroscience* 20(4), 580–91.

van der Zee, E. and Slack, J. (eds.) (2003). *Representing Direction in Language and Space.* Oxford: Oxford University Press.

Vorwerg, C. (2001). *Raumrelationen in Wahrnehmung und Sprache. Kategorisierungsprozesse bei der Benennung visueller Richtungsrelationen* [Spatial relations in perception and language. Categorization processes in referring to visual directional relations]. Wiesbaden: Deutscher Universitätsverlag.

—— and Kronhardt, J. (2008). *Saliency Impact on Initial Frame-of-reference Selection, and Consistency in Verbal Localization*. Manuscript submitted for publication.

—— and Rickheit, G. (1998). 'Typicality effects in the categorization of spatial relations', in C. Freksa, C. Habel, and K. F. Wender (eds.), *Spatial Cognition. An Interdisciplinary Approach to Representing and Processing Spatial Knowledge*. Berlin: Springer.

—— —— (1999). 'Richtungsausdrücke und Heckenbildung beim sprachlichen Lokalisieren von Objekten im visuellen Raum [Direction terms and hedges in the verbal localization of objects in visual space]', *Linguistische Berichte* 178: 152–204.

Vorwerg, C. and Rickheit, G. (2000). ‘Repräsentation und sprachliche Enkodierung räumlicher Relationen [Representation and linguistic encoding of spatial relations]’, in C. Habel and C. von Stutterheim (eds.), *Räumliche Konzepte und sprachliche Strukturen*. Tübingen: Niemeyer.

—— and Weiß, P. (2008). *How Verb Semantics Affects the Interpretation of Spatial Prepositions*. Manuscript submitted for publication.

Wagner, S. M., Nusbaum, H., and Goldin-Meadow, S. (2004). ‘Probing the mental representation of gesture: is handwaving special?’, *Journal of Memory and Language* 50: 395–407.

Wahlster, W. (ed.) (2006). *SmartKom: Foundations of Multimodal Dialogue Systems*. Berlin: Springer.

Walker, M. A. (1992). ‘Redundancy in collaborative dialogue’, in *Proceedings of COLING 92*.

Watson, M., Pickering, M. J., and Branigan, H. P. (2004). ‘Alignment of reference frames in dialogue’, in *Proceedings of the 26th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

—— —— ——. ‘Why dialogue methods are important for investigating spatial language’.

Weiß, P., Grabowski, J., and Miller, G. A. (1996). ‘Factors affecting spatial deictic communication: a comparison of German and American English’, in *Proceedings of the Second Colloquium on Deixis ‘Time, Space and Identity’, 28–30 March 1996, Nancy*. Nancy: Centre de Recherche en Informatique.

Werner, S., Krieg-Brückner, B., and Herrmann, T. (2000). ‘Modelling navigational knowledge by route graphs’, in C. Freksa, W. Brauer, C. Habel, and K. F. Wender (eds.), *Spatial Cognition II: Integrating Abstract Theories, Empirical Studies, Formal Methods, and Practical Applications*. Springer.

Wraga, M., Creem, S. H., and Profitt, D. R. (2000). ‘Updating displays after imagined object and viewer rotations’, *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26: 151–68.

Wunderlich, D. (1981). ‘Linguistic strategies’, in F. Coulmas (ed.), *A Festschrift for Native Speaker*. The Hague: Mouton.

—— and Herweg, M. (1991). ‘Lokale und Direktionale [Locatives and directionals]’, in A. von Stechow and D. Wunderlich (eds.), *Semantik: Ein internationales Handbuch der zeitgenössischen Forschung*. Berlin/New York: de Gruyter.

Wyatt, J. (2005). ‘Planning clarification questions to resolve ambiguous references to objects’, in *Proceedings of the Fourth Workshop on Knowledge and Reasoning in Practical Dialogue Systems*. IJCAI’05, Edinburgh.

Zacks, J. M., Mires, J., Tversky, B., and Hazeltine, E. (2000). ‘Mental spatial transformations of objects and perspective’, *Spatial Cognition and Computation* 2: 315–32.

—— Rypma, B., Gabrieli, J., Tversky, B., and Glover, G. H. (1999). ‘Imagined transformations of bodies: an fMRI investigation’, *Neuropsychologia* 37(9): 1029–40.

—— Tversky, B., and Iyer, G. (2001). ‘Perceiving, remembering, and communicating structure in events’, *Journal of Experimental Psychology: General* 136: 29–58.

Ziegler, F., Mitchell, P., and Currie, G. (2005). ‘How does narrative cue children’s perspective taking?’, *Developmental Psychology* 41: 115–23.

Zwaan, R. A. and Radvansky, G. A. (1998). ‘Situation models in language comprehension and memory’, *Psychological Bulletin* 123: 162–85.

Name Index

Subject Index

Page numbers in *italics* refer to figures, illustrations or photographs.